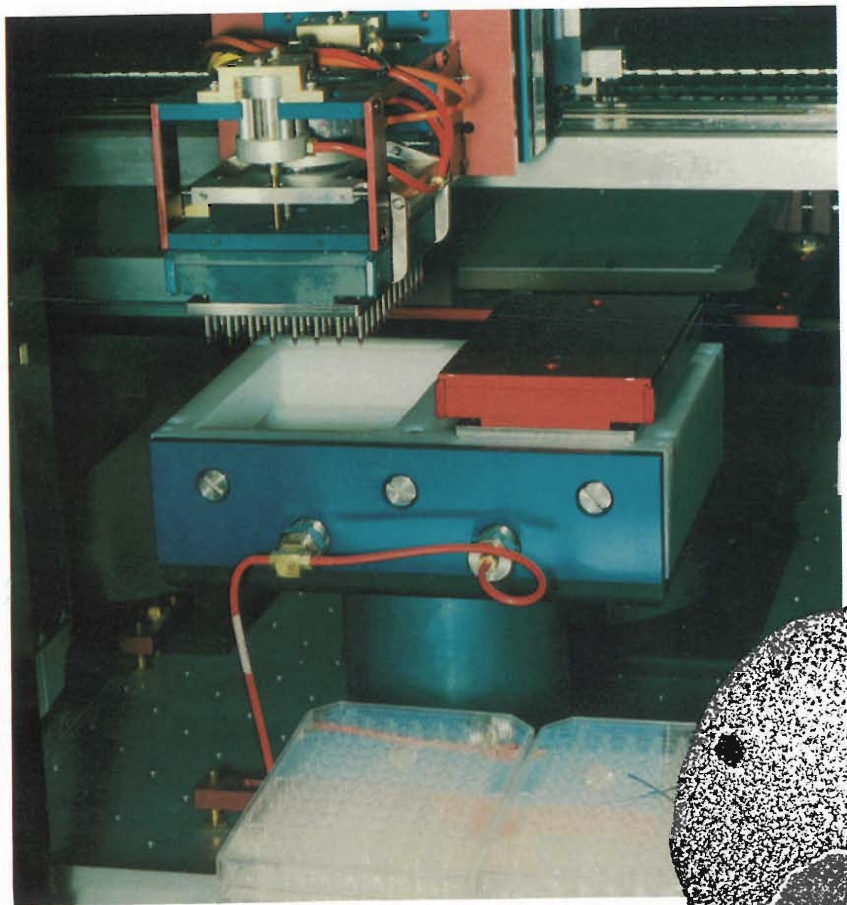
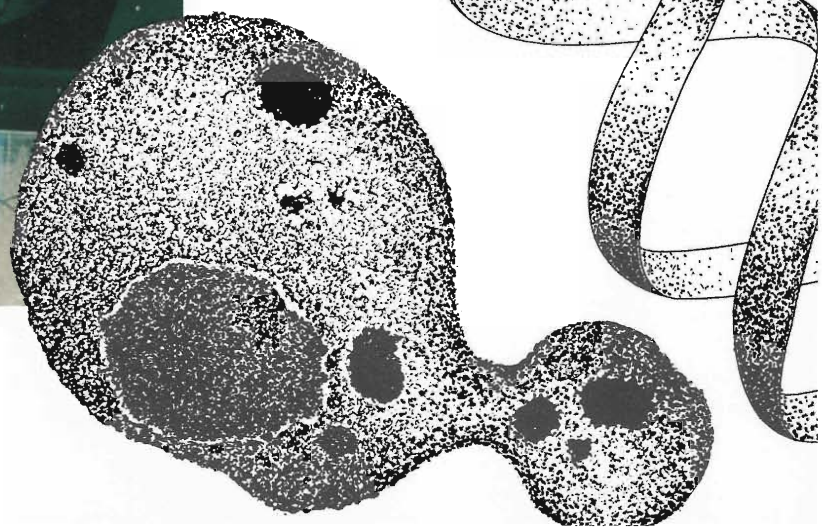
recombinant clones
for mapping and sequencing

DNA



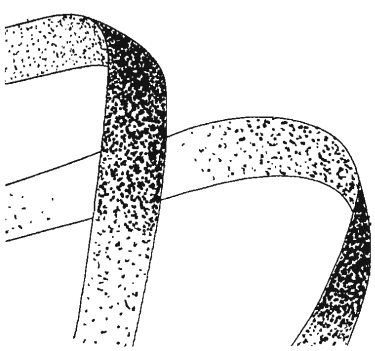


LIBRARIES

Larry L. Deaven

Until the 1970s it was nearly impossible to isolate and purify single genes in sufficient quantity for biochemical analysis and DNA-sequence determination. The difficulty was largely due to the small size of many genes (2000 to 10,000 base pairs, or 2 to 10 kbp) and the large size of complex genomes such as the human genome (3 billion base pairs). In order to obtain 1 milligram of a 2-kbp human gene, such as the β-globin gene, all of the DNA in all of the cells of twenty-four people would have to be used as the starting material. Even if it were practical to obtain that much DNA, the problem of separating the DNA sequences that encode β-globin from the rest of the DNA would be very difficult. A solution to this problem was found during the recombinant-DNA revolution through the development of a technique called molecular cloning. By using molecular-cloning techniques, a small fragment of DNA can be duplicated, or amplified, into an unlimited number of copies.

Shown on these pages are the two common host cells for molecular cloning, the bacterium E. coli *and the yeast* S. cerevisiae*; a popular cloning vector, the λ phage, with its icosahedral head and long tail; a membrane containing a gridded array of recombinant clones to which DNA probes have been hybridized; and a robotic device developed at the Laboratory that creates those gridded arrays.*



Molecular cloning of a gene requires three ingredients: one copy or a few copies of the gene to be cloned, a biological cloning vector, and a host cell. Cloning vectors are small molecules of DNA, often circular, that can be replicated within a host cell. Host cells are usually single-celled organisms such as bacteria and yeast. The first step of the cloning process is to combine the DNA fragment containing the gene sequence with the DNA of the cloning vector. If the vector DNA is circular, the circle is cut and the gene to be cloned is joined to each end of the opened circle. The new, somewhat larger circle of DNA is called a recombinant molecule, as is any molecule formed from a cloning vector and an inserted DNA fragment. The recombinant molecule can now be allowed to enter a host cell, where it is duplicated by the replication machinery of the host cell. Each time the recombinant molecule is replicated a new copy of the gene it contains is produced. Furthermore, each of the two daughter cells formed by the division of the original host cell receives copies of the recombinant molecule. When the host cell has grown into a colony, it is referred to as a recombinant clone, and the DNA fragment contained within each cell of the colony is said to have been cloned.

If we apply the cloning process to the production of 1 milligram of the human β-globin gene, a few copies of the gene would be inserted into plasmid cloning vectors. (Plasmids are small circular DNA molecules found in bacteria.) The recombinant plasmids would then be added to *E. coli* bacterial cells. Some of the cells would be entered, or "transformed," by a recombinant plasmid and would begin to produce copies of the cloned β-globin gene. Using this approach, just 2 liters of nutrient solution would produce enough *E. coli* cells to yield 1 milligram, or many trillions of copies, of the β-globin gene. Molecular cloning removed the barriers that had prevented the biochemical and molecular analysis of individual genes in complex genomes.

During the recombinant-DNA revolution of the 1970s molecular cloning was also applied to the study of entire genomes with even more dramatic results. In that application, instead of cloning one gene at a time, all of the DNA in a genome is cut into small fragments and each of those fragments is cloned. The resulting collection of cloned fragments is called a DNA library. The word "library" was chosen because collectively, those cloned fragments contain all of the genetic information in an organism. Like a library of reference books, a library of cloned human DNA, for example, represents a collection of reference material for studying the genetic information in human beings. However, whereas conventional libraries are ordered collections of information, DNA libraries are unordered and uncharacterized collections of recombinant clones. Those collections provide the starting materials for almost all the current techniques used to decipher the instructions contained in DNA.

Two general features of libraries make them a remarkable resource. First, individual clones from a library can easily be isolated from the other clones. If the host cells are bacteria, a small portion of the library can be placed on a culture dish where each bacterium will form a colony of identical cells. Each colony can then be transferred to an individual culture dish and grown into a large population. Since each population contains a different cloned DNA insert, any region of the genome can be made accessible for analysis and sequencing. Second, a DNA library is a renewable resource. The clones can be grown individually or collectively to replace any portions of the library that have been consumed. Therefore, a library is, in a sense, permanent. It can be repeatedly used or shared with other laboratories with little or no depletion of the original recombinant clones.

Just as there are legal libraries and medical libraries and scientific libraries, there are various types of DNA libraries. Each type is classified according to the vector used in library construction and the source of insert DNA. For example, DNA libraries constructed from the DNA in human cells are called human-genomic libraries. Ideally they contain all of the DNA sequences present in the human genome. Human cDNA libraries contain only those sequences utilized in protein coding. They are constructed by isolating messenger RNA (mRNA) molecules from human tissue and converting them into complementary DNA (cDNA) by the action of the enzyme reverse transcriptase. The cDNA fragments are then cloned. Because mRNA molecules are derived from the protein-coding portions of genes (see "Protein Synthesis" in "Understanding Inheritance"), cDNA libraries contain the sequences within genes that are expressed as proteins. Thus a brain cDNA library would be made from the mRNAs in brain cells and would contain only those DNA sequences expressed as brain proteins. Similarly, a liver cDNA library would contain those DNA sequences whose expression as protein is necessary to the proper functioning of liver cells.

A library is further classified according to the vector used in its construction. Since different vectors tend to carry DNA inserts with a limited range of lengths, classification by cloning vector, in effect, specifies the average length of the inserts within the recombinant clones of the library. Each type of library offers particular advantages for particular applications.

A primary goal of the Human Genome Project is to construct a physical map of each human chromosome. A physical map of a chromosome is an ordered collection of clones selected from one or more DNA libraries. Collectively, those clones carry inserts that include all of the DNA in the chromosome, and through the mapping process each cloned insert is ordered according to its position along the length of the chromosome (see "Physical Mapping" in "Mapping the Genome"). Thus the construction of a physical map is somewhat analogous to the cataloging of documents in a conventional library. Many of the recent improvements in cloning technology have been due to the initiation of the Human Genome Project and specifically to the need for physical maps of each human chromosome. Libraries with large DNA inserts make the mapping process both faster and easier, so considerable attention has been given to the development of cloning systems that can faithfully maintain and propagate large DNA inserts.

Although much effort is directed to ordering, or physical mapping, of the clones in a library, unorganized libraries are also useful tools. Through a process called library screening, cloned fragments of DNA that contain a sequence of interest can be retrieved from a library. The sequence of interest might be a region of a chromosome that contains a gene or some other genetic landmark. To find a particular clone, the library is screened with a DNA probe whose sequence is identical to a small portion of the region of interest. The probes may be synthesized, but often they are obtained from small-insert genomic-DNA libraries or from cDNA libraries. For example, a labeled probe containing a unique DNA sequence from the hemoglobin gene can be used to identify the clones in a DNA library whose inserts contain all or a portion of that gene.

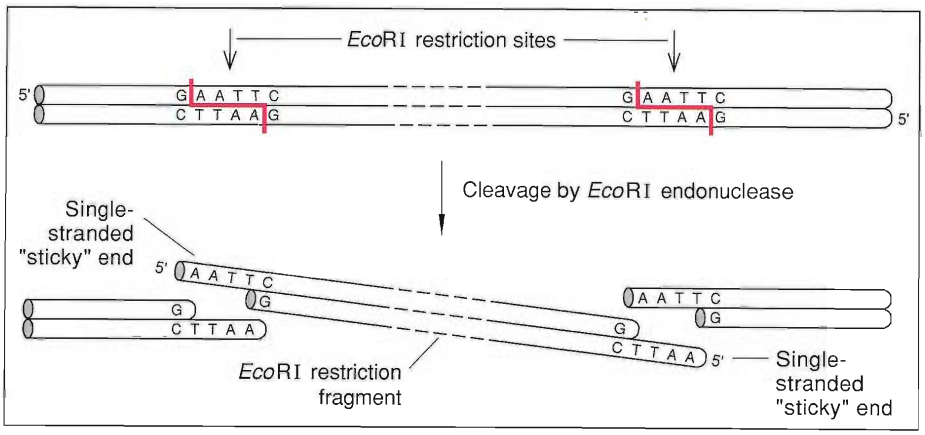


Figure 1. Restriction-Enzyme Cleavage

A restriction enzyme cleaves DNA at each recognition site on the DNA molecule. In this illustration the restriction enzyme is *Eco*RI, which cleaves DNA having the sequence 5′-GAATTC. Both strands of the DNA are cleaved between the G and A bases, leaving a tail with the sequence TTAA on each cut end. The tails are complementary to each other, so two pieces of DNA that have both been cleaved with *Eco*RI can be joined end-to-end to make a recombinant molecule. If a DNA molecule has tails that facilitate joining, it is said to have sticky ends.

DNA libraries are vital to much of the research in molecular genetics and to most of the activities sponsored by the Human Genome Project, including the construction of physical maps, the sequencing of DNA fragments, the isolation of genes, and the search for polymorphic genetic-linkage markers. Details of those activities are discussed elsewhere in this issue.

This article points out applications of DNA libraries but focuses primarily on the libraries themselves. It includes a brief history of the discoveries that led to the first DNA libraries and descriptions of the various types of libraries, their construction and manipulation, and the pioneering work here at Los Alamos National Laboratory on the construction of human-chromosome-specific DNA libraries.

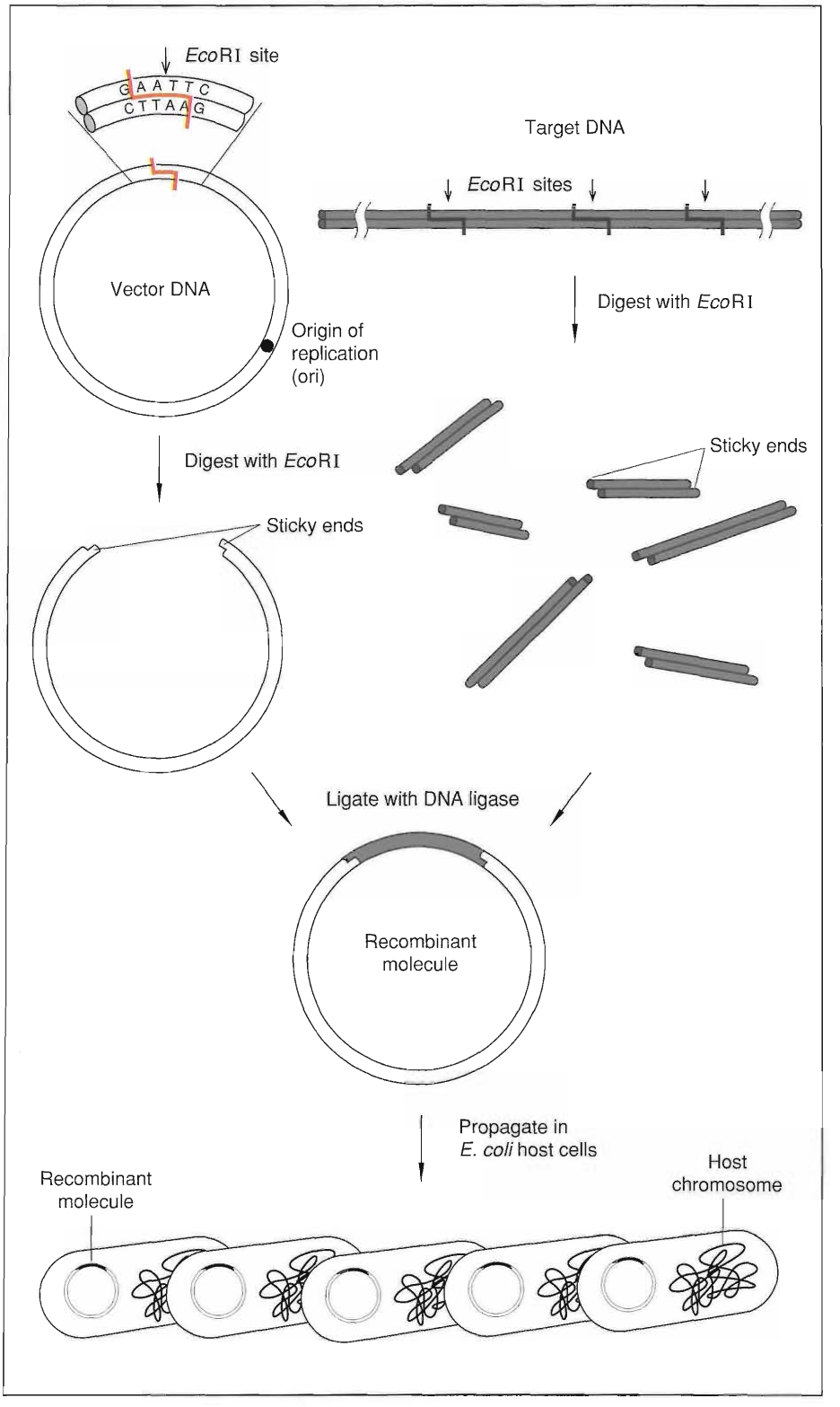
Historical Background

The ability to construct libraries of recombinant clones depends on a very long series of discoveries and technological developments in DNA biochemistry. These include the discovery that DNA is the carrier of genetic information in 1944, the determination that DNA has a double-helical structure in 1953, and the unraveling of the genetic code in the 1960s. However, the first essential step in the origin of recombinant-DNA technology was the discovery in 1970 of a group of bacterial enzymes now called class-II restriction endonucleases, or simply restriction enzymes. Those enzymes help to protect the bacterium from the DNA of invading viruses by recognizing certain specific sequences in DNA and cleaving the viral genome within or near those recognition sites. (The bacteria produce other enzymes, called methyltransferases, that prevent the restriction enzymes from cutting the bacteria's own DNA.)

Figure 1 shows how the restriction enzyme *Eco*RI cuts double-stranded DNA into fragments. *Eco*RI is called a six-base cutter because it recognizes

Figure 2. Construction and Propagation of Recombinant Molecules

A circular DNA molecule containing an origin of replication, which allows the replication machinery of *E. coli* to reproduce the molecule, is used as a vector to carry foreign DNA into a host cell. The DNA of the circular vector and linear molecules of target DNA are digested or cut with the same restriction enzyme (*Eco*RI). The result is linear vector molecules and fragments of target DNA, which all have complementary "sticky" ends that permit the molecules to be joined by DNA ligase. When the vector and insert are joined, the resulting recombinant DNA molecule is inserted into an *E. coli* cell. Billions of copies of the recombinant molecule are made as the transformed cell replicates through many generations to form a bacterial colony. Each copy contains the short fragment of human DNA that was inserted into the original recombinant molecule.



the six-base DNA sequence 5′-GAATTC and can cut DNA molecules at every site where that sequence occurs.

Like the recognition sequences of most restriction enzymes, that of *Eco*RI is a "palindrome," meaning that the sequence on one strand is identical to the complementary sequence on the other strand when both are read in the 5′-to-3′ direction. *Eco*RI cuts the phosphodiester bond between the G and the A nucleotides on both strands. Thus the enzyme produces a staggered cut so that the two cut ends have single-stranded tails, or so-called sticky ends. Those ends are useful for making recombinant molecules because any two fragments generated by the same restriction enzyme have identical sticky ends and therefore can be held together by hydrogen bonding. The two fragments can then be permanently joined, or recombined, by enzymes called DNA ligases.

Restriction enzymes provide a tool for cutting DNA in a reproducible way and they produce fragments that can easily be joined to other similarly cut fragments. Moreover, the many different restriction enzymes make it possible to cut large molecules of DNA into fragments of a controlled average size. In addition to six-base cutters, there are four-base and eight-base cutters. If the four bases A, T, G, and C were distributed randomly in DNA molecules, on average a given four-base sequence would occur every 256 base pairs, a given six-base sequence approximately every 4 kbp, and a given eight-base sequence approximately every 66 kbp. In actual practice, restriction-enzyme cleavage sites, or restriction sites, do not occur at random. For example, since the enzyme *Not*I recognizes the eight-base sequence 5′-GCGGCCGC, it would be expected to produce fragments averaging 66 kbp in length after all available sites are cut. But when *Not*I is allowed to completely digest human DNA, that is, to cut all its restriction sites in a sample from the human genome, it produces fragments that have an average length of 1 million nucleotides because its recognition sequence is rarer than expected (in particular the sequence 5′-CG is rare in mammalian genomes). Nevertheless, by selecting the proper restriction enzyme, it is possible to repeatably cut DNA molecules into fragments of different average lengths. Fragment size may also be adjusted by allowing the enzyme to cut only a portion of the available restriction sites. For example, if *Eco*RI is permitted to cut only one 5′-GAATTC sequence in five, the resulting average fragment size will be 20 kbp rather than 4 kbp. This "partial digestion" is accomplished by using a shorter incubation period or a lower concentration of enzyme than a complete digest would require. The ability to reduce large molecules of DNA to smaller fragments of controlled average size is a critical step in the construction of libraries because most cloning vectors accept only DNA inserts whose lengths fall within a limited range.

Just two years after the discovery of restriction enzymes, the first experiments were performed that created recombinant DNA molecules. DNA containing genes from a bacterium and from a bacterial virus was inserted into the genome of simian virus 40, a virus that infects mammalian cells. The two types of DNA were initially in the form of closed loops. The restriction enzyme *Eco*RI was used to cut the loops and the resulting linearized molecules were joined to form recombinant molecules. The ultimate objective of that work was to use the simian virus as a biological vector to carry foreign genes into mammalian cells and to see if the foreign genes would be expressed in their new environment. Concerns over the potential hazards of recombinant molecules halted research on gene transfer into mammalian cells for several years; nevertheless, the experiment clearly demonstrated that a restriction enzyme would cut DNA in a predictable manner and that restriction fragments from two different organisms could be joined.

Shortly after that experiment, molecular cloning techniques were extended and improved. In one set of experiments, a plasmid containing a single *Eco*RI restriction site as well as a gene for resistance to the antibiotic tetracycline was purified from *E. coli* and a method was devised for introducing the plasmid into other *E. coli* cells that were not resistant to tetracycline. The transformed cells were then grown on agar (a culture medium) mixed with tetracycline. Some of the bacteria grew into colonies, demonstrating that they had taken up the plasmid and that it was functioning. Following that experiment, the plasmid was recombined with a second plasmid containing a gene for resistance to the antibiotic kanamycin. The recombinant plasmids also transformed host cells and conferred antibiotic resistance. Finally, experiments demonstrated that DNA from two different species could be recombined and propagated as a recombinant plasmid. A gene encoding a ribosomal RNA in the toad *Xenopus laevis* was recombined with *E. coli* plasmid DNA and propagated in *E. coli* host cells. The general approach for those experiments is shown in Figure 2.

As more experience was gained in recombinant-DNA technology, new cloning vectors were developed and methods for growing and handling recombinant molecules were further improved. The possibility of cloning fragments of DNA that represented all of the genetic information in the human genome began to look achievable. An intermediate step in this direction was the construction of a recombinant-DNA

library from the DNA in fruit flies (*Drosophila melanogaster*) in 1974. Several recombinant plasmids containing either unique-sequence or repetitive-sequence inserts were isolated from the library and localized or mapped to specific regions of the *Drosophila* chromosomes. That work suggested the potential value of constructing libraries from the total DNA in a complex organism, selecting clones from the library that contain genes of interest, and then using those clones to find the chromosomal location of the cloned gene sequences. Further progress led to the construction of a library of DNA from human embryonic liver tissue in 1978 and the selection of clones from the library that contained human α- and β-globin genes. That experiment clearly demonstrated that DNA libraries could provide a starting point for mapping the human genome and for studying gene structure and expression.

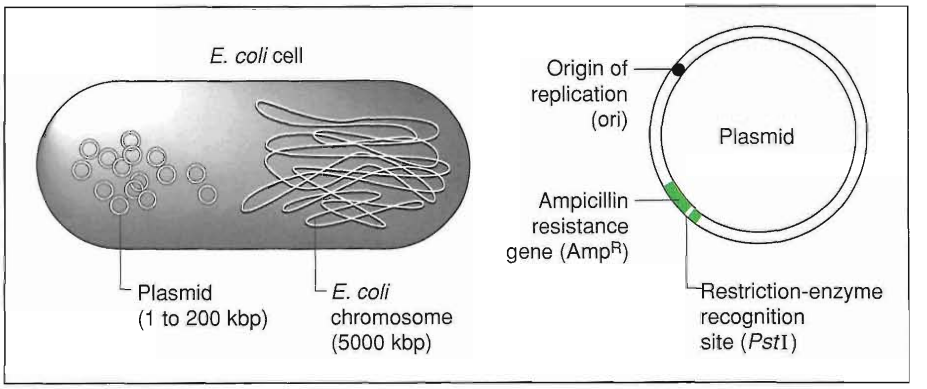


Figure 3. Plasmids

Plasmids are small, circular DNA molecules that occur naturally in *E. coli* and other bacteria. They all contain a replicon, a DNA sequence that enables the host bacterium to replicate them. The replicon includes an origin of replication (ori). Many contain restriction-enzyme cleavage sites and DNA sequences that encode antibiotic-resistance genes. For instance, the plasmid shown here contains an ampicillin-resistance gene and a single cleavage site for the restriction enzyme *Pst* I. These natural properties were exploited to adapt plasmids for use as the first vector systems.

Library Construction

Continued progress since 1978 has made it possible to construct many kinds of recombinant-DNA libraries. Libraries differ in the preparation of the target DNA, the choice of the host strain, and the design of the cloning vector. Each variation produces a library that has advantages for specific applications. The most significant characteristics of a library are usually determined by the choice of the cloning vector, so a description of the vectors currently in use provides a convenient way of defining the variety of libraries.

This article focuses primarily on the four types of vectors currently used in constructing libraries for the physical mapping of complex genomes. In common use today are plasmids, bacteriophage (phage) genomes, cosmids, and yeast artificial chromosomes (YACs). The first three vectors are all referred to as *E. coli*-based systems because they are propagated within the common intestinal bacterium *Escherichia coli*. The fourth is called a yeast-based system because propagation occurs within the baker's yeast *Saccharomyces cerevisiae*.

Whereas the original versions of some of these vectors were similar to wild-type organisms, the modern vectors are highly engineered for a variety of uses. For example, DNA sequences not essential for replication have been removed from some wild-type vectors to provide space for large DNA inserts. Molecular biologists have also inserted sequences into the vectors that help in incorporating and manipulating DNA inserts and in recovering the inserts from recombinant clones.

E. coli-based Cloning

Plasmids. As mentioned above, plasmids were the first vectors to be used in constructing recombinant clones. These small chromosomes are often found in *E. coli* cells along with the main bacterial chromosome. Plasmids are circular, double-stranded DNA molecules that range in length from 1 to 200 kbp and are thus considerably smaller than the main chromosome, which is about 5 million base pairs long (see Figure 3).

Plasmids frequently contain genes that are advantageous to the bacterial host. Among these are genes that confer resistance to antibiotics and genes that produce restriction enzymes. Every plasmid also includes DNA sequences called replicons, each of which contains an origin of replication and the other elements the plasmid needs in order to be replicated by bacteria. Although some types of plasmids replicate only when the main chromosome replicates and tend to exist as a single copy within the host cell, most plasmids commonly used as cloning vectors replicate independently of the main chromosome and exist in multiple copies, from ten to five hundred, within the host. The entry of a plasmid, whether engineered or natural, into a

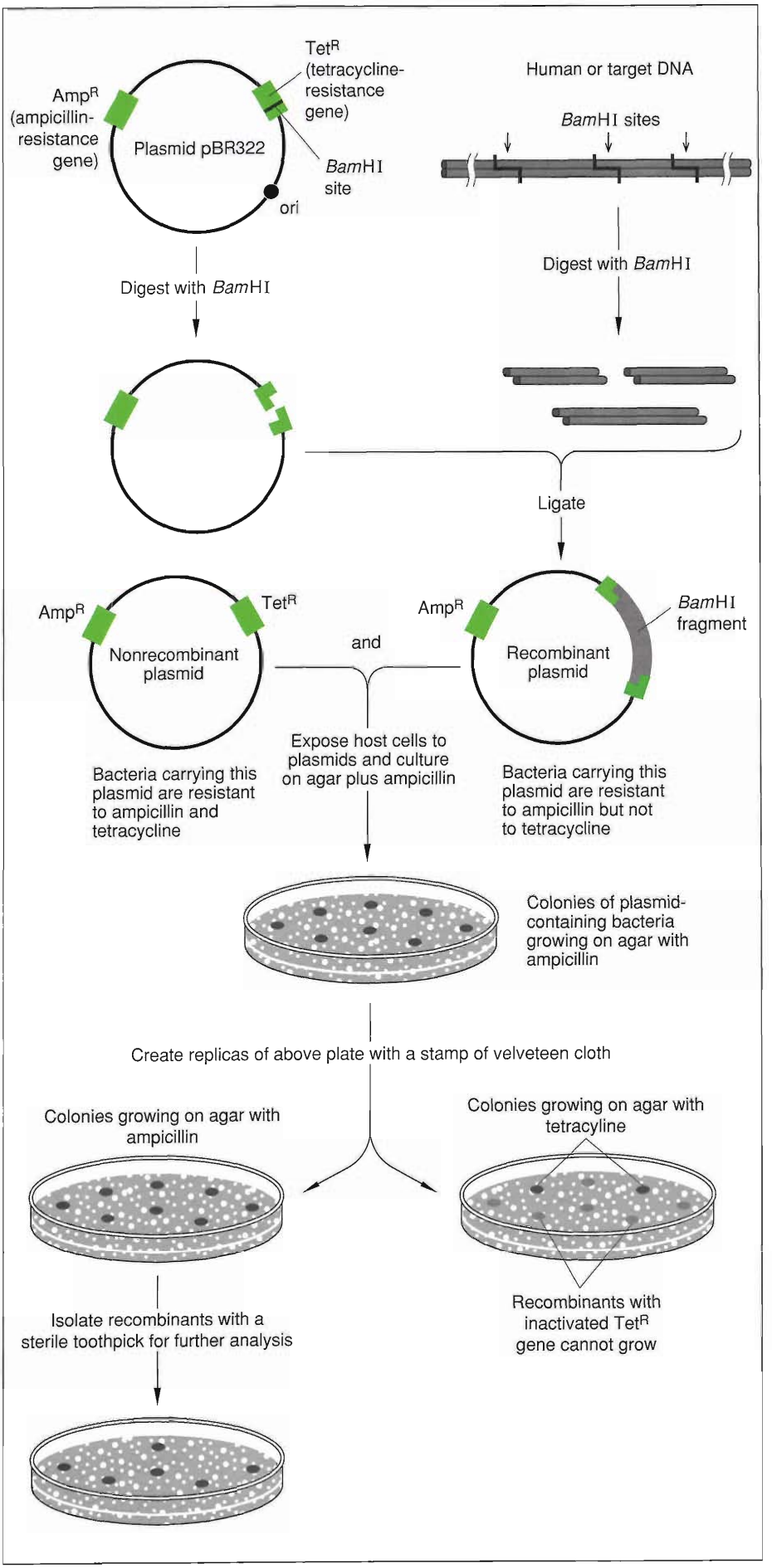


Figure 4. Selection of Recombinant Clones

The plasmid vector pBR322, containing genes for ampicillin resistance and tetracycline resistance, allows clones containing foreign-DNA inserts to be distinguished from clones lacking inserts. Digestion with *Bam*HI opens the circular molecule at a point within the tetracycline resistance gene. A target DNA fragment produced by *Bam*HI digestion can then be inserted to make a circular recombinant plasmid. The presence of the insert inside the tetracycline gene inactivates the gene. Thus nonrecombinant plasmids provide resistance to both ampicillin and tetracycline, whereas recombinant plasmids provide resistance only to ampicillin. A population of plasmid-free host cells is exposed to the plasmids and then spread on culture dishes containing agar mixed with ampicillin. Only host cells that were transformed by either recombinant or nonrecombinant plasmids multiply and form clones in the presence of ampicillin. A portion of each clone is transferred to each of two other dishes in a way that preserves the relative positions of the clones. One dish has ampicillin in the agar; the other has tetracycline. The recombinant clones are those that grow in ampicillin and do not grow in tetracycline. This selection technique, called insertional inactivation, was used in the early plasmid vectors. Now it is more common to use a single antibiotic resistance gene as a selectable marker and select transformed cells directly on the basis of response to the appropriate antibiotic. The formation of nonrecombinant plasmids is suppressed by chemical techniques (such as removing the phosphate groups from the ends of the vector so that the ends can not bind to each other).

bacterial cell is called transformation. The exact mechanism of entry is unknown, but all the methods for increasing the frequency of transformation involve increasing the permeability of the pores in the bacterial membrane, which presumably allows the plasmid to pass through. Even when those methods are employed, only a small fraction (1 in 10,000) of the cells in a bacterial population are stably transformed when exposed to a solution containing plasmids.

The naturally occurring plasmids used as cloning vectors in the 1970s contained features that could be exploited both in the cloning process and in the process of selecting or identifying clones containing recombinant plasmids. Figure 4 illustrates such a plasmid. It contains a single cutting site for the restriction enzyme *Bam*HI. The restriction site is located in the middle of a gene conferring resistance to the antibiotic tetracycline. To form a recombinant plasmid, the vector is cut at that site with *Bam*HI and then ligated with a fragment of target DNA that was also produced by digestion with *Bam*HI. The DNA insert thus separates the antibiotic-resistance gene into two pieces and inactivates the gene. If a bacterial cell is transformed by that recombinant plasmid, that host cell will be sensitive to tetracycline. A bacterial cell transformed by a plasmid with no insert will be resistant to tetracycline. Thus the tetracycline-resistance gene not only contains a cloning site, or site for insertion of a foreign DNA fragment, but also acts as a selectable marker to differentiate recombinant clones containing DNA inserts from clones containing plasmid vectors but no foreign DNA insert.

The plasmid vectors developed in the early 1970s were useful, but they had many limitations. They replicated poorly, had a limited number of selectable markers, and contained restriction sites for at best two restriction enzymes. The plasmids used today have been engineered to overcome these limitations. Some plasmids even contain regulatory regions that facilitate the expression of foreign genes contained within the DNA insert and genes that can change the color of a bacterial colony and thus allow visual identification of clones containing recombinant plasmids.

However, plasmids have two limitations that cannot be overcome. First, plasmids are inefficient at transforming bacteria. Second, plasmids containing long DNA inserts are particularly inefficient at transformation, and tend to lose portions of the inserts as they are replicated. Therefore, plasmids are usually used to carry short inserts on the order of 4 kbp in length. To clone all the DNA in the haploid human genome (3 billion base pairs) would require 750,000 plasmids each containing a different DNA insert. To find a particular gene in such a library, all of those clones would have to be screened. Those limitations spurred the development of new cloning vectors with higher transformation efficiencies and the ability to accommodate larger inserts. The first of the new vectors was the genome of a bacterial virus, bacteriophage λ.

Bacteriophage λ. Bacteriophage, or phage, are viruses that infect bacteria. Being extremely simple biological systems, they had been extensively studied since the 1930s. In the 1970s they were seen as promising cloning vectors because their DNA genomes are readily replicated by the cellular machinery of the host bacterium and because, unlike plasmids, they have a natural and efficient mechanism of entry into a bacterial host.

An intact λ phage has a protein coat consisting of an icosahedral head and a rod-like tail. The head contains the phage genome, a double-stranded linear DNA molecule about 48 kbp in length, with short, complementary single-stranded ends of 12 nucleotides each. Those cohesive ends are called *cos* sites. During replication, the phage tail attaches to a bacterial host cell, and the phage genome enters the host's interior. There the DNA molecule may incorporate itself into the bacterial chromosome. Alternatively, in the "lytic" life cycle used in cloning, the DNA cyclizes by base pairing of the *cos* sites and begins to express genes involved in the replication of phage DNA. Initially, the replication process forms a long strand of DNA that consists of hundreds of copies of the λ-phage genome. Such a strand is called a concatamer. Then the phage DNA directs the synthesis of proteins for the head and tail as well as enzymes that cut the concatamer into individual λ genomes and package each one into a phage head. When the cell contains between 100 and 200 new phage particles (about 20 minutes after infection), λ proteins cause it to rupture, or "lyse," and the released phage particles infect surrounding cells (see Figure 5).

A phage particle added to a monolayer, or "lawn," of bacterial cells growing on an agar plate produces through that infection cycle a clear area called a plaque containing lysed bacterial cells and replicated phage. A visible plaque contains a population of from 1 to 10 million identical phage particles.

A section in the middle of the λ-phage genome contains a cluster of genes that are unnecessary for its replication in *E. coli* cells. To make the λ-phage genome into a vector, either the DNA is cut or that middle section of DNA is removed, leaving the left and right end fragments (called arms) that are essential for λ replication. The arms are then attached to an insert. If the insert is not too different in size from the DNA that it

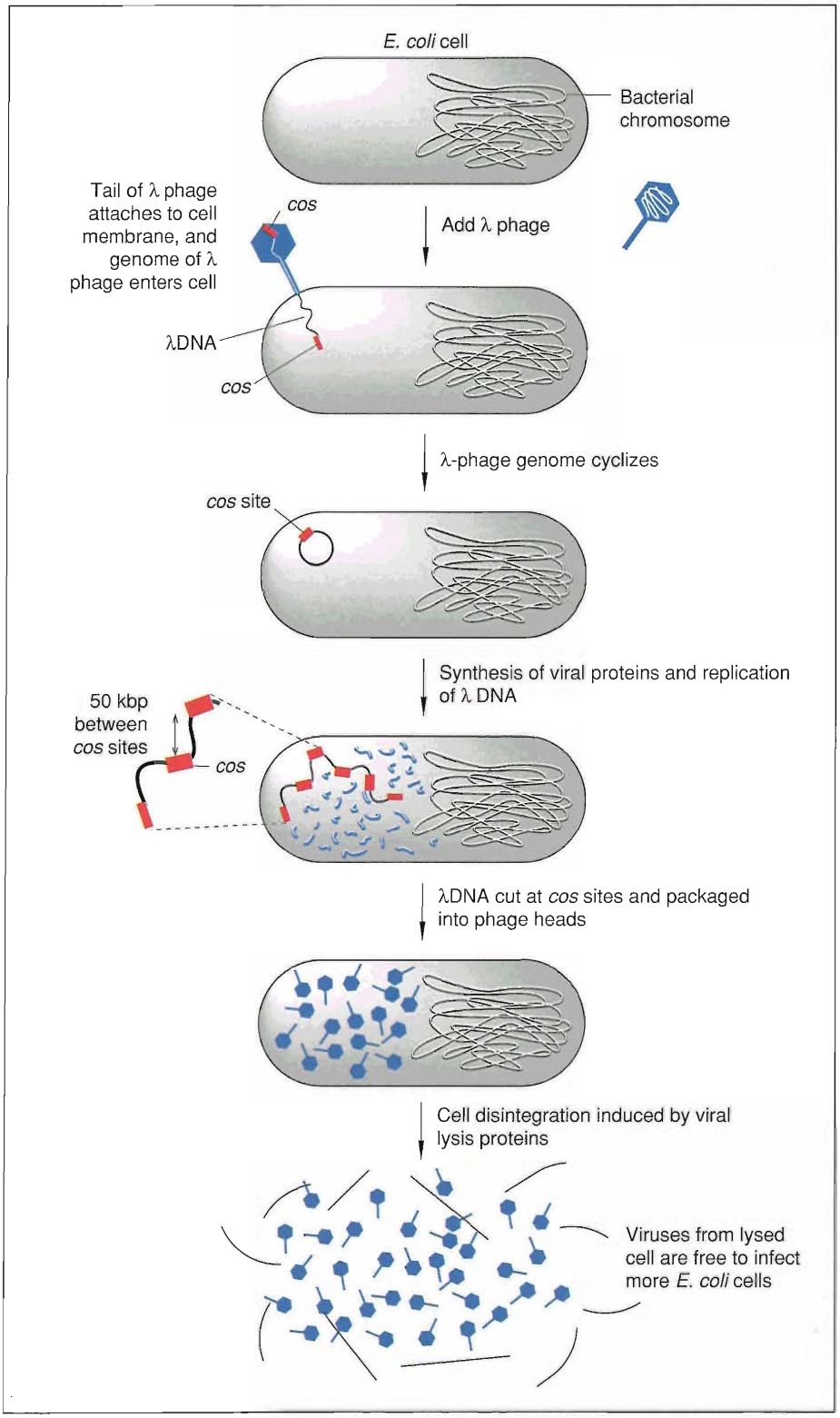


Figure 5. Lambda-Phage Lytic Life Cycle

The tail of a phage particle attaches to the surface of a host *E. coli* cell and the phage genome enters the cell. The genome cyclizes by base pairing of the complementary, single-stranded "terminus" on each end (the cohesive ends or *cos* sites). The viral DNA then directs the synthesis of proteins necessary for its replication and enzymes and structural proteins necessary for the assembly of phage particles. The product of genome replication is a long chain, or "concatamer," of many copies of the viral chromosome joined end to end at the *cos* sites. When 100–200 copies of the viral DNA have been made, the concatamer of DNA is cleaved at the *cos* sites into individual phage chromosomes by phage enzymes that recognize and cut them. Phage enzymes then package these genomes into phage particles that are released by cell lysis and can infect new bacterial cells.

replaces, the resulting recombinant DNA molecule can be packaged to make an infective λ particle.

Packaging is performed in vitro using "packaging extracts" that include the enzymes and structural proteins needed for head and tail assembly. Packaging extracts are isolated from two strains of *E. coli* engineered from natural strains whose chromosomes contain λ DNA (as a result of the phage's nonlytic life cycle). Under appropriate growth conditions each of the engineered strains makes some of the proteins necessary to package λ-phage particles. If a single strain produced all the packaging proteins, phage coats would form inside the cell and the cells could not be used as a source of packaging extracts. Therefore in each strain the λ DNA has a mutation that prevents the bacterium from producing one protein essential for assembly of λ-phage particles. The strains differ in which protein is missing. The unassembled phage proteins are extracted from cells of each strain and combined in vitro with each other and the recombinant DNA so that the DNA can be packaged into phage particles. Once the phage particles have been produced in a test tube, the phage infection cycle described above will, in a very short time, generate recombinant phage clones containing millions of copies of the insert DNA.

As suggested above, for the packaging to work the size of the insert must be similar to the size of the phage DNA it replaces (or if no phage DNA was removed, the insert must be small). In practice, inserts usually range from 12 to 22 kbp in length. With inserts of 20 kbp the DNA in the human genome could be fragmented and included in a library made up of 150,000 recombinant λ-phage particles, a considerable improvement over the 750,000 plasmids that would be required. Phage also have the advantage of transforming hosts far more efficiently than plasmids do; typically one phage particle in ten infects a host bacterium.

A cloning vector based on λ phage was first used in 1974. Since that time many versatile and sophisticated vectors have been derived from the wild-type phage (see Figure 6). This progress is due in large part to the extensive studies of the genetics and physiology of bacteriophage beginning in the late 1930s and continuing today. Without that accumulation of detailed knowledge, the use of λ phage as a central tool of molecular biology would have been delayed and might well not have been developed at all.

Cosmids. Bacteriophage-λ vectors made it possible to construct libraries with inserts of up to 22 kbp. However, many genes contain on the order of 35 to 40 kbp. In order to clone those genes as single inserts, a vector with greater capacity was needed. In 1979 the first account was published of a successful library based on a more capacious vector called a cosmid.

Cosmids were engineered to combine desirable features of plasmids and λ phage (including the *cos* site, whence the word "cosmid"). Phage can transform hosts efficiently and maintain long inserts without deletions. Nevertheless, the size of their inserts is limited because a λ-phage head can hold no more than 52 kbp of DNA, and the phage requires 30 to 40 kbp for replication and packaging. On the other hand, plasmid vectors need only a replicon for reproduction in host cells, a drug-resistance gene for use as a selectable marker, and restriction sites for inserting foreign DNA. A cosmid is designed to reproduce as a plasmid and be packaged as a λ phage. It contains all the necessary ingredients for reproduction in *E. coli* and for other cloning functions and is only 5 to 6 kbp long. Cosmids can therefore accept inserts as large as 47 kbp and still be packaged in a phage protein coat to facilitate entry into host cells.

This synthetic vector is grown as a plasmid in *E. coli* and then isolated. To prepare the circular vector for cloning, it is cleaved by a restriction enzyme to produce a linear molecule containing a *cos* site. Next a DNA insert is ligated with the vector. The ligation produces long concatamers in which inserts alternate with vectors. When phage packaging extracts are added to the concatamers, the *cos* site in each vector is cleaved, producing individual phage chromosomes. Chromosomes in the appropriate size range are packaged into phage particles that can infect bacteria. Once inside the host cell, the recombinant DNA cyclizes and reproduces as a plasmid. Because inserts in cosmids have an average size of about 40 kbp, a cosmid library containing all of the DNA in the human genome would require approximately 75,000 clones, about half as many as a λ-phage library would require. Unfortunately, some cosmids, if not maintained under optimal conditions, may lose portions of their inserts during replication.

Yeast-based Cloning

Yeast Artificial Chromosomes. Cosmid vectors fulfilled some of the needs for longer cloned inserts. However, during the 1980s new genes were discovered that are too large to be cloned as single fragments in cosmids, and attempts to map large segments of the human genome were hindered by the small size of the inserts in λ and cosmid libraries. A new cloning system that accommodates longer inserts was first reported in 1987. The recombinant molecules are called yeast artificial chromosomes (YACs) because they are maintained and reproduced as chromosomes in yeast

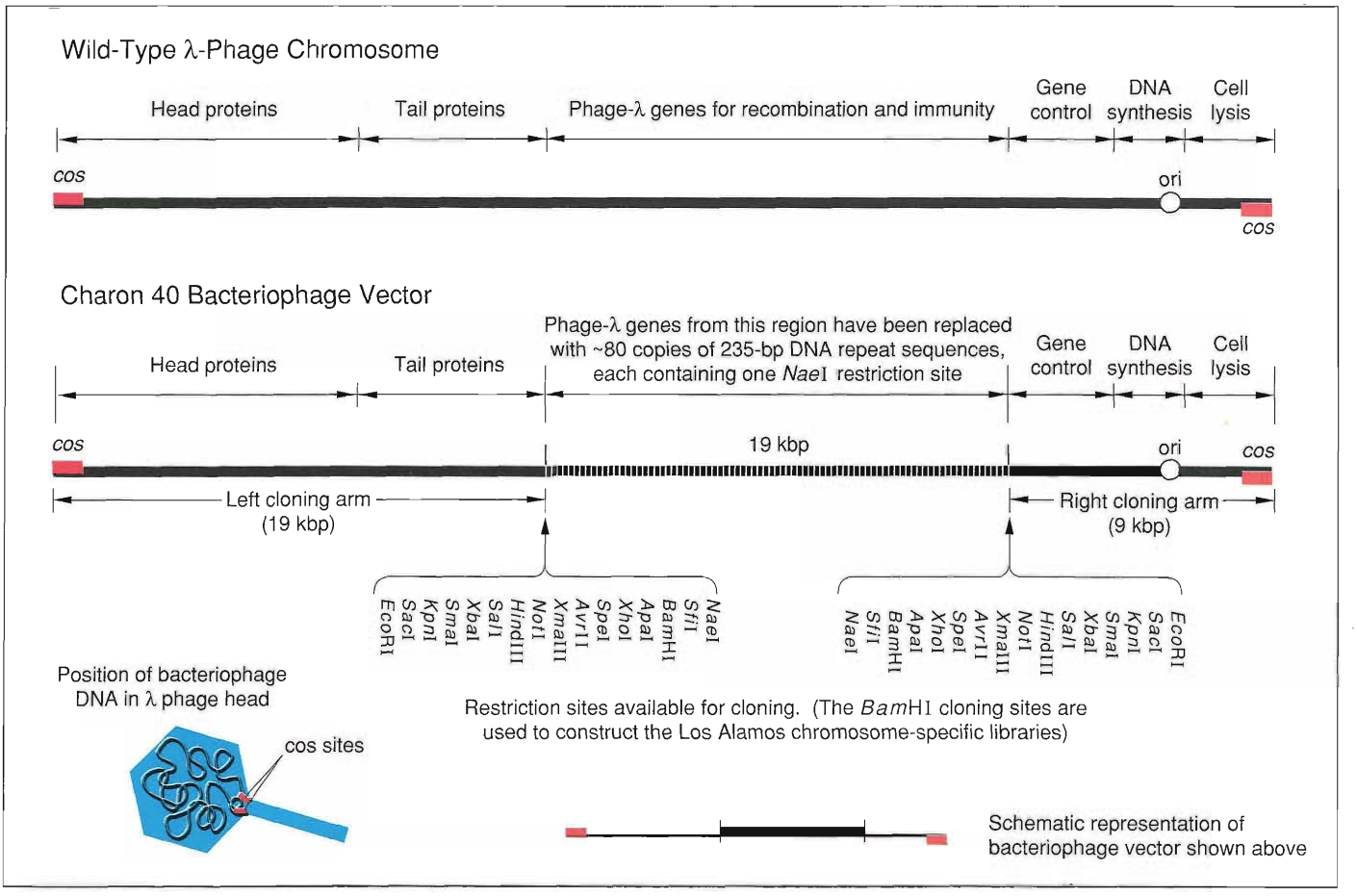


Figure 6. Engineering the Genome of Wild-Type λ Phage into a λ-Phage Vector

The genome of the wild-type lambda phage is divided into six regions according to the locations of genes that encode various functions. In the vector Charon 40, a region that is not necessary for replication is replaced with 80 copies of a 235-base-pair sequence that contains a cleavage site for the restriction enzyme *Nae*I. When Charon 40 is cut with *Nae*I, the repeat-sequence region is reduced to small fragments that can be separated from the cloning arms of the vector by gel electrophoresis. The section of Charon 40 on each side of the repeat-sequence region (enlarged) contains a single restriction site for each of a number of restriction enzymes. These sites have been added to Charon 40 to increase its versatility as a cloning vector.

host cells. The vector arms contain a yeast centromere, two yeast telomeres, and a yeast origin of replication, the elements necessary for yeast cells to replicate the recombinant molecules in the same way they replicate yeast chromosomes.

YAC vectors, like cosmid vectors, are highly engineered and are produced as plasmids in *E. coli*. The first YAC vectors were single plasmids containing all the yeast sequences listed above as well as a plasmid replicon, one or more markers to use in selecting *E. coli* cells containing YAC vectors, and two restriction sites for the same enzyme: one between the sequences that give rise to the telomeres and one at which to insert target DNA. Cloning with these YAC vectors is similar in approach to λ-phage and cosmid cloning. The vector is cleaved at the insertion site and between the telomere sequences. The cleavage produces two vector arms that are ligated with the insert to produce a YAC. The YAC is then allowed to transform a yeast cell; once inside the host cell, it behaves as a stable chromosome.

The YAC vector carries a gene that suppresses the host strain's production of a red pigment. The commonly used cloning site is within that suppressor gene. If nonrecombinant YAC vectors

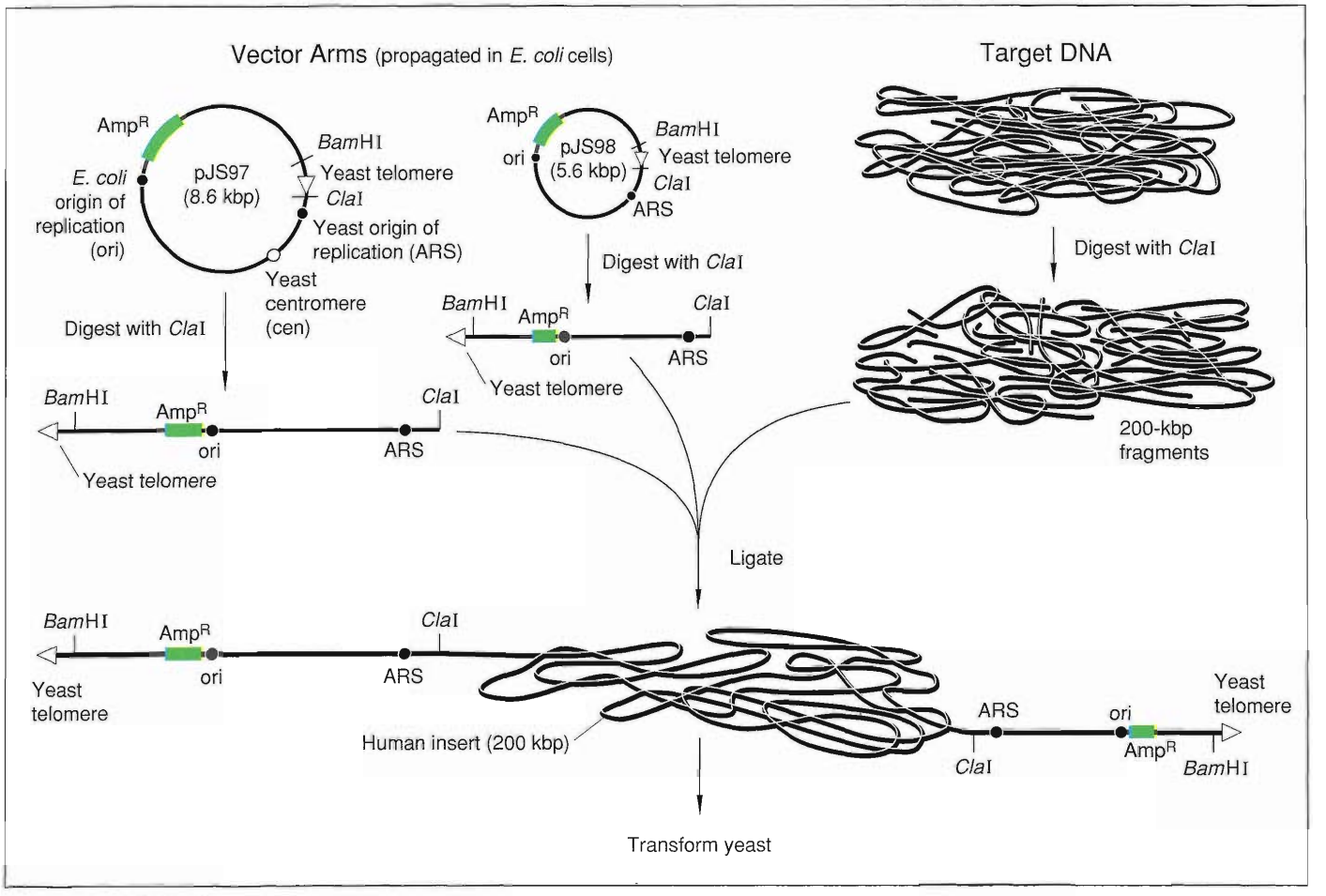


Figure 7. YAC Cloning

The two arms of the YAC shown are manufactured separately in *E. coli* as plasmids pJS97 and pJS98. Each contains an ampicillin-resistance gene, an *E. coli* replicon (including an origin of replication) so that it can propagate as a plasmid, a yeast origin of replication (labeled ARS), a yeast telomere, and several restriction sites, including one for *Cla*I located at the end of the telomere. Only pJS97 contains a yeast centromere and a pigment-suppressor gene that changes the color of yeast colonies. The plasmids are linearized and the target DNA is fragmented, both by cutting with *Cla*I. The vector arms thus produced each have a yeast telomere sequence at one end (arrow) and a *Cla*I tail at the other end. The fragments of target (human) DNA are then ligated to the vector arms to form a YAC that can transform yeast cells. Promoters for T7 RNA polymerase (not shown) are located near the *Cla*I restriction site. These sequences are used in generating RNA probes from the ends of the insert.

transform yeast cells, the resulting yeast colonies are white. Insertion of target DNA inactivates that gene, causing the formation of red rather than white yeast colonies and providing a rapid means of identifying the colonies that contain the target DNA.

In 1991 a new type of YAC vector was reported that has additional advantages (see Figure 7). This vector is produced as two separate plasmids: one carries the centromere and serves as one arm of the YAC, and the other serves as the second YAC arm. Each arm has a selectable marker to identify transformed hosts, and the arm containing the centromere also contains a pigment-suppressor gene used to monitor the number of YACs in each host cell and their stability against deletions. Again the host strain produces red pigment. When one YAC is present in each cell, the colony is pink; the presence of two YACs in each cell causes the colony to be white; and an unstable colony (one in which some cells are losing the YAC as they divide) has red and pink sectors.

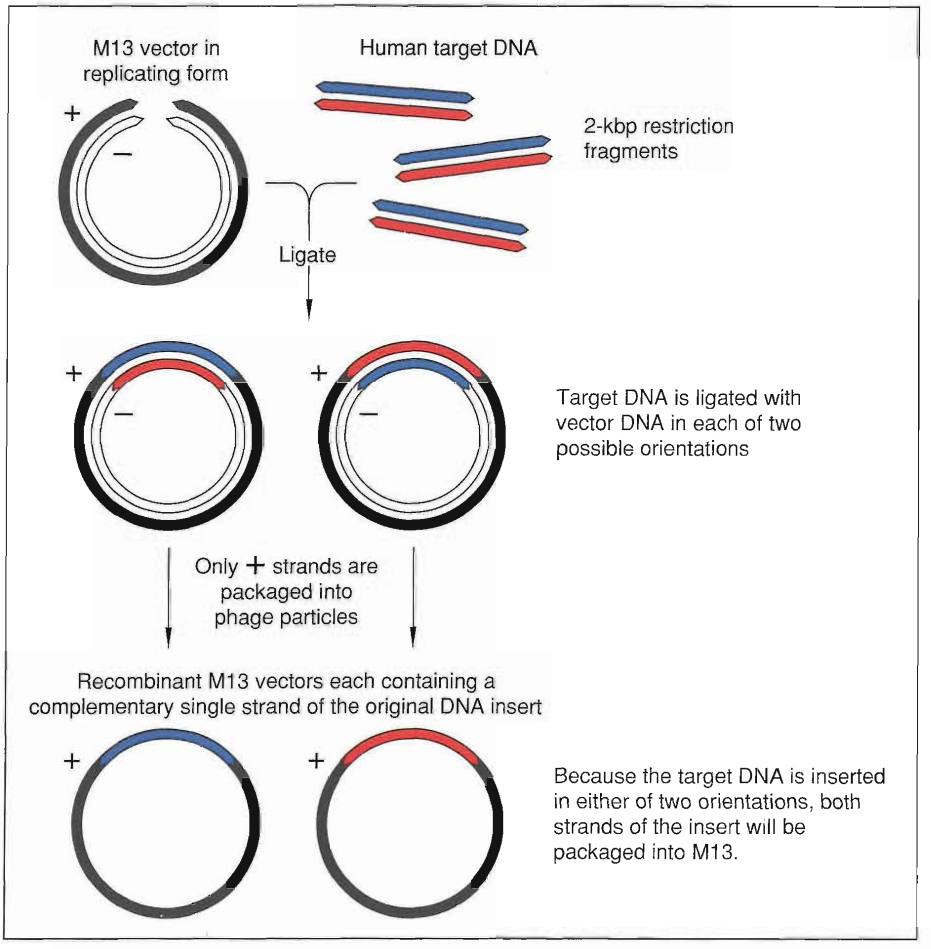


Figure 8. Cloning in Bacteriophage M13

The double-stranded circular replicating form of the phage genome is recovered from infected *E. coli* cells and then cut with a restriction enzyme. The resulting vector is ligated to target DNA cut with the same enzyme. The recombinant molecules are allowed to enter bacteria as plasmids do. Inside the bacterium, they replicate themselves and produce phage proteins. Only the + strands of the phage genome are packaged as viral progeny. Nevertheless, roughly equal amounts of both strands of the insert are obtained because some inserts enter phage DNA molecules oriented so that one strand of the insert is incorporated into the phage's + strand, and some inserts enter so that the complementary strand is incorporated into the + strand. Finally the + strands are packaged into phage particles and leave the host cell (without damaging it).

To allow the generation of probes from the ends of the DNA inserts, each arm contains a promoter sequence from the T7 phage that is located near the sites of attachment to the insert. RNA polymerase from phage T7 can bind to the promoters and transcribe the ends of a DNA insert into RNA. These RNA "end probes" are useful in characterizing the YAC inserts. For example, one can try to hybridize either end probe from one YAC to all of the other YACs in a library and thus locate overlapping YACs for chromosome walking.

The YAC cloning system has the advantage of being able to maintain and propagate inserts up to a million base pairs long. The average insert size in many YAC libraries is 200 to 400 kbp, five to ten times the average size in a cosmid library, and the human genome may be covered in 7500 YAC clones.

A major disadvantage of the YAC cloning system is that it often produces chimeras. Chimeras are YACs whose inserts are composed of more than one piece of target DNA. For example, a chimeric YAC with a 400-kbp human insert may contain 300 kbp from human chromosome 3 and 100 kbp from human chromosome 10. Chimeras complicate the construction of physical maps of overlapping clones, and they must be identified to avoid mapping errors. Unfortunately, identifying chimeras is laborious. Moreover, since 40 to 60 percent of the clones in most YAC libraries are chimeric, chimeras add a significant amount of work to the mapping process. Techniques developed at the Laboratory for producing nonchimeric YACs are described in "Libraries from Flow-sorted Chromosomes."

Bacteriophage M13. Another vector used to construct libraries is derived from the filamentous *E. coli* phage M13. M13 clones are particularly convenient for DNA sequencing. An M13 phage particle consists of single-stranded DNA packaged into a narrow cylindrical protein coat. The strand of DNA in the phage particle is designated as the + strand. When the phage infects *E. coli* cells, the DNA replicates to form about 300 double-stranded (+/-) copies, but only the + strands from those copies are packaged into progeny virus particles. Figure 8 shows the method of cloning with M13. The double-stranded form is used as the cloning vector. Small fragments (about 2 kbp) are inserted into any of several restriction

sites engineered into the M13 phage vector. The recombinant molecules enter the host cells as plasmids and replicate as phage. The + strands produced by the replication are used as a template for Sanger dideoxy sequencing (see "DNA Sequencing" in "Mapping the Genome"). Because the sequence at the M13 insertion site has been determined, an oligonucleotide (short DNA sequence) can be synthesized to serve as a universal primer for the dideoxy sequencing of any DNA fragment that is cloned into M13.

Host Cells

Host cells are as important in cloning as vectors. To work well in cloning, host cells should be as accessible as possible to the introduction of the vector, they should facilitate library screening, and they should alter inserts as rarely as possible. In addition, while host organisms should provide conditions for robust growth of recombinant molecules, they must also be sufficiently disabled to have no significant probability of surviving outside of laboratories. The need for safe cloning systems was a major concern for scientists in the early years of recombinant-DNA research, and progress was delayed until host bacterial strains were developed that had many features to prevent the escape of transformed cells from laboratories. For example, the weakened strains require chemicals not likely to be found in nature and have cell walls that burst in the presence of low salt concentrations or a trace of detergent. In hindsight many of the concerns about unexpected, hazardous properties of recombinant organisms have turned out to be unwarranted. Nevertheless, the early guidelines and regulations helped reassure the public that recombinant-DNA procedures would not result in new diseases or the spread of bacterial antibiotic resistance.

Once the issue of safety was appropriately addressed, the development of host-vector systems accelerated, in large part because of the wealth of information available on the genetics and biochemistry of *E. coli* cells. Some vectors are so specialized that they can be propagated only in a single host strain. Others can grow in a wide variety of strains so that a host strain can be selected according to the requirements of a specific cloning application. In general, strains of bacteria that produce restriction enzymes are avoided because those strains do not propagate inserts that contain a susceptible cleavage site. Some strains of bacteria produce enzymes called methyltransferases that add methyl groups to certain bases in DNA. Those enzymes protect bacteria from their own restriction enzymes by altering the structure of recognition sites. Strains with active methyltransferase genes are also unsuitable for cloning because they would produce recombinant DNA molecules that could not be cleaved by certain restriction enzymes and therefore could not be used in experiments involving those enzymes. For bacteriophage-λ vectors, host strains must be susceptible to λ infection. For plasmid cloning, strains free of nonvector plasmids must be used in order to recover the recombinant molecules without contamination by other plasmids. Other interactions between vectors and host cells can serve to identify and isolate cells that contain vectors. For example, a host strain that requires a particular amino acid can be used with a vector that contains a gene for the production of the amino acid. When grown on a medium lacking the amino acid, only bacteria that have incorporated the vector will survive.

Wild-type *E. coli* produce enzymes that recombine DNA strands containing homologous sequences. Because human DNA contains many sequences that are repeated in various places in the genome, there was considerable concern that recombinant inserts would be rearranged and deleted when propagated in *E. coli*. From the beginning strains of *E. coli* deficient in recombination enzymes were used in cloning. Now many such strains have been engineered, and reports of DNA rearrangements are far outnumbered by studies that find no rearrangements after extensive propagation of recombinant molecules. There are, unfortunately, a few types of sequences that are known to replicate poorly or not at all in the *E. coli* environment, primarily repetitive sequences such as the DNA in the centromeric region of a chromosome.

As might be expected of the relatively new YAC-*S. cerevisiae* system, the choice of host strains is limited; in fact, only two are available. The desired features are similar to those for *E. coli* strains: ease of transformation, stable maintenance of artificial chromosomes, and compatibility with various selection and recovery systems. The two yeast host strains in widespread use differ primarily in the selectable markers they contain, and because YACs can be readily transferred from one strain to the other, the features of the two host strains are complementary. A useful addition to the available yeast host strains would be a strain deficient in recombination pathways, which would reduce the incidence of chimeric inserts in YACs.

Genomic Libraries Constructed from Cellular DNA

As mentioned previously, the first libraries to contain DNA inserts from total cellular DNA were constructed using phage vectors. Those early

libraries, like many used today, were designed to facilitate the isolation of genes and nearby regulatory sequences for studies of gene structure and function. Because a single gene might span several fragments in the library, a library designed to be searched for genes should consist of overlapping cloned fragments. Then a series of cloned fragments that overlap each other and span the entire gene can be identified. Overlapping fragments are made by partially digesting with a restriction enzyme the DNA extracted from many cells. In partial digestion of target DNA, the restriction enzyme cleaves a random subset of the restriction sites in each of the many copies of the target molecules and thereby produces a population of overlapping fragments, which can be used in constructing a library of overlapping clones.

To identify clones in the library that duplicate all or part of a gene of interest, the library must be screened with a gene probe, a single-stranded short segment of DNA or RNA composed of a sequence complementary to the sequence of the gene. The first probes were cDNAs constructed from specialized cells that produce large amounts of specific mRNAs. Later, it became possible to construct cDNA libraries that contained sequences complementary to most of the mRNAs found in specific tissues, such as brain tissue.

As illustrated in Figure 9, the first step in screening a bacteriophage library with a gene probe is to grow a lawn of *E. coli* host cells in a set of agar-coated Petri dishes. About 150 to 200 recombinant phage particles from the library are added to each plate. When plaques have formed, a filter membrane is placed on each agar surface. Some phage particles from each plaque adhere to the membrane, so a pattern of invisible spots identical to the pattern of plaques on the Petri plate is formed on

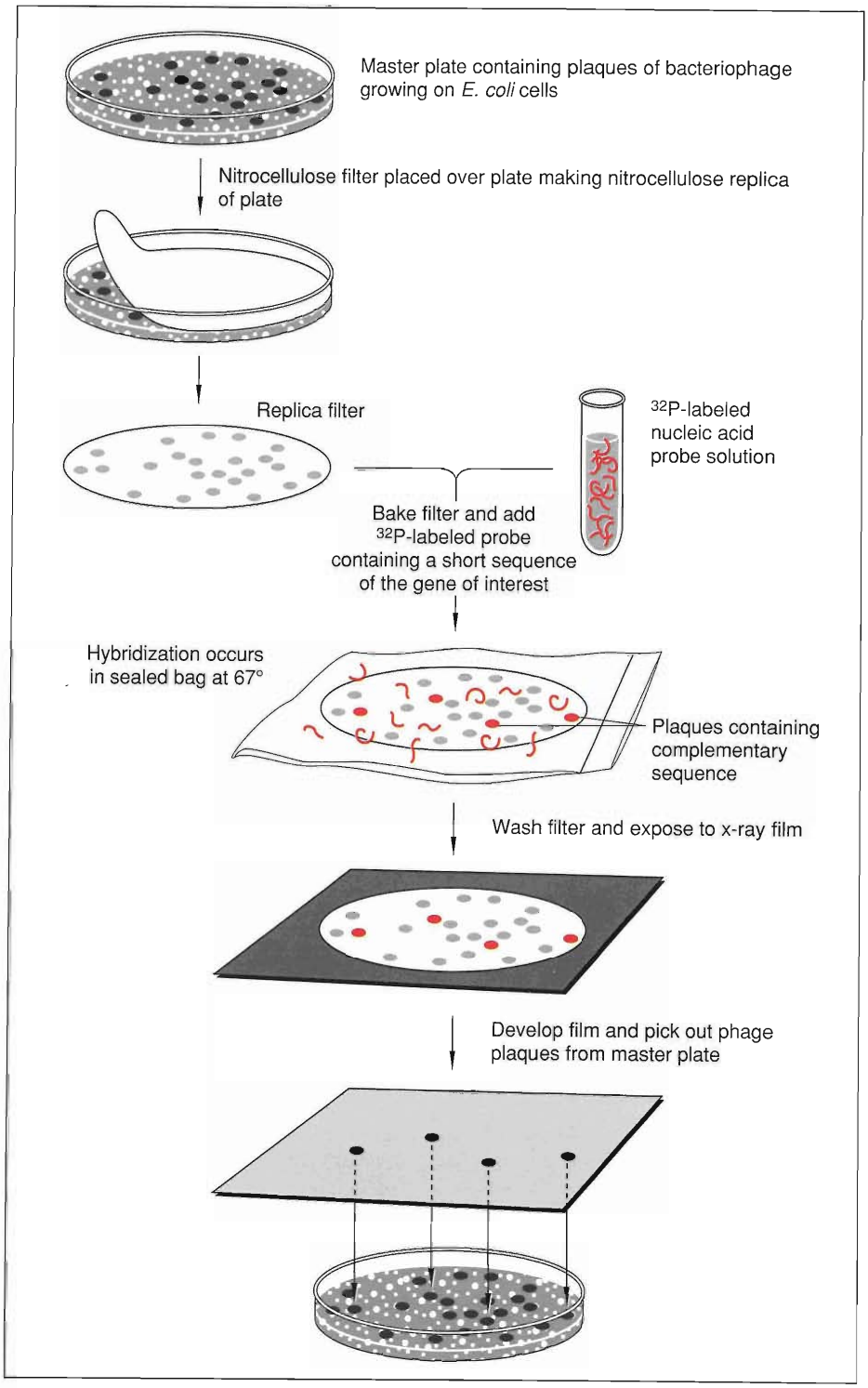


Figure 9. Screening a Phage Library

Recombinant phage from the library are allowed to form plaques on a set of master plates. A nitrocellulose filter membrane is placed on the surface of the agar. Some phage adhere to the membrane, providing a copy of the plaque arrangement. The filter is first treated with sodium hydroxide to lyse the phage particles and denature the DNA they contain. The filter is then baked to prevent the DNA from renaturing and to fix it in position. The membrane is exposed to a radioactively labeled probe, which hybridizes only to those spots containing a DNA sequence complementary to the probe sequence. Then the radioactive spots are detected by making a sandwich consisting of a sheet of x-ray film and the filter enclosed in a plastic wrap. When the x-ray film is developed, spots appear at the same positions as the positions of the plaques in the Petri dish that hybridized to the probe.

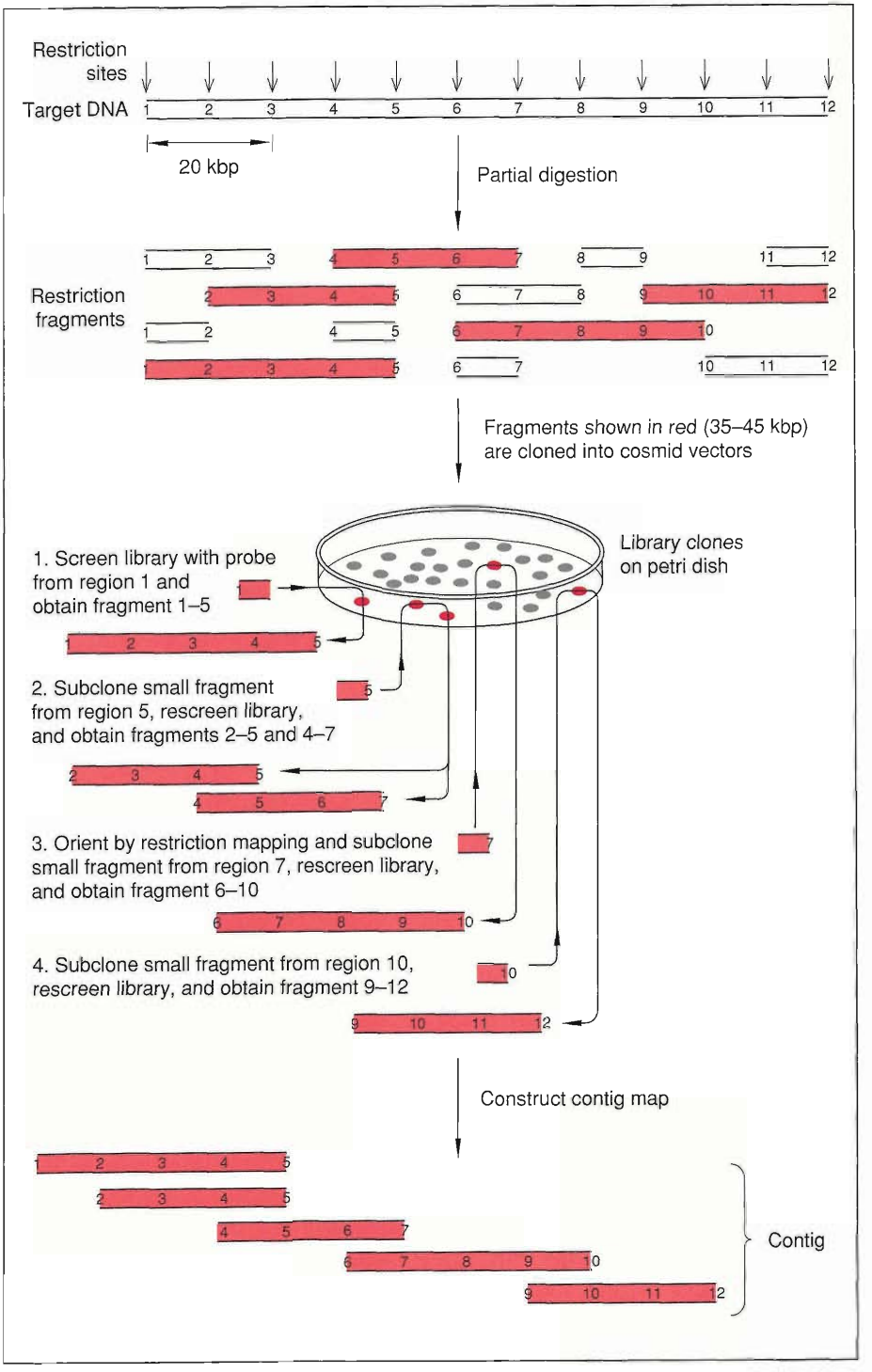


Figure 10. Chromosome Walking

The aim of this technique is to recover fragments of cloned DNA that span a gene or any DNA region of interest. A DNA library constructed from partially digested DNA provides a source of overlapping cloned fragments of the entire genome. A DNA probe known to be close to the gene of interest is used to screen the library for fragments that contain sequences complementary to that of the probe. In the illustration, the probe comes from region 1 and the screening shows that fragment 1–5 contains that region. Now a small single-copy portion of the DNA in region 5 is used as a probe to identify other clones in the library that contain DNA from region 5. In our example, this second screening identifies two fragments: 2–5 and 4–7. A third screening using a probe from region 7 identifies the fragment 6–10. The process is repeated until clones containing the entire region of interest have been identified.

the filter membrane. Each spot is composed of phage particles from a single plaque. A radioactively labeled probe is then allowed to hybridize to the DNA on the membrane. Any spots on the membrane that contain DNA complementary to the probe become radioactively labeled. When a labeled spot is identified, the plaque corresponding to that spot can be located in the Petri dish. Each such plaque contains all or part of the gene (or of a member of the gene family) of interest. The phage in those plaques can then be isolated and regrown to provide more DNA for further study of the gene.

If no single clone in the library of overlapping clones includes the entire gene, a process called chromosome walking is used to identify a series of overlapping clones whose inserts span the gene sequence (see Figure 10). In this technique, the clone identified by the initial gene probe is cut into smaller fragments with one or more restriction enzymes. A short segment of single-copy DNA from one end of the clone insert is then used as a probe to re-screen the library and identify an overlapping clone. A probe from the endmost fragment of the second clone is then generated and used to find a third clone that overlaps the second. The process continues until the set of overlapping clones spans the entire gene. Chromosome walking thus produces a contig map of the region containing the gene (see "Physical Mapping" in "Mapping the Genome"). Sometimes a segment of DNA within the gene is not present in the library, in which case several libraries must be used to complete the walk. Each step in chromosome walking takes a few weeks to a month. If a phage library is used, as many as several hundred thousand phage plaques must be screened with each probe, and the distance covered with each step in the walking process

may be only 1 to 1.5 kbp. Obviously the use of libriaries with large inserts is advantageous. The workload is reduced by a factor of about two if a cosmid library is used, and by a further factor of five to ten if a YAC library is available. Here again cloning systems capable of carrying larger and larger inserts are desirable.

Another kind of library made from total cellular DNA is a library of the DNA in a "monochromosomal" hybrid-cell line. Hybrid-cell lines are constructed by inducing cells from two different species to fuse together to become one cell. Hybrids are commonly made by fusing human cells with mouse or Chinese hamster cells. Initially, the hybrid cell contains two complete sets of chromosomes, one from the human cell and one from the rodent cell. However, each time the hybrid cell divides, it tends to lose some of its human chromosomes. Some of the hybrid cells lose all of their human chromosomes, whereas others retain one or more human chromosomes for various lengths of time. A single cell from a population of hybrid cells can be grown into a clone of identical cells and analyzed for the specific human chromosomes it contains. An alternative method for selecting a hybrid cell containing a specific human chromosome involves growing the population of hybrid cells in a special culture medium that selectively kills cells lacking the desired chromosome. These techniques have been used to create a series of cell lines called somatic-cell hybrid panels in which each cell line contains only one copy of one human chromosome in a rodent background.

Libraries made from such cell lines have advantages over libraries made from normal human cells. Because DNA from only one human chromosome is present in the target DNA, all human DNA inserts in the library are known to come from that chromosome. Furthermore, hybrid-cell libraries contain inserts from only one copy of the human chromosome, whereas in libraries made from human cells containing chromosome pairs there is no easy way to determine whether a clone from the library originated from one or the other member of the homologous pair. On the other hand, libraries made from hybrid cells have the disadvantage that the clones containing human inserts may constitute as little as a few percent of the total number of clones, which makes selecting the human inserts from the rodent background very laborious.

Libraries from total cellular DNA have been made in phage, cosmid, and YAC vectors. The use of phage libraries gradually gave way to cosmid libraries, especially for studies of genes too large to be cloned and propagated in λ or plasmid vectors, and for studies that required assembling sets of overlapping clones that spanned large regions of DNA. Cosmids, in turn, were replaced by YACs, and current interests are in improving YAC technology or developing alternate cloning systems that can carry very large inserts. Any of these libraries may be screened for clones of interest. The purpose of screening may be to construct physical maps for small regions around genes or gene families, to construct maps for entire chromosomes, or to select polymorphic markers for use in genetic-linkage mapping (see "Modern Linkage Mapping" in "Mapping the Genome"). Inserts in cosmid or phage clones may also be subcloned into M13 vectors for DNA sequencing.

Library Amplification and Storage

Amplification of a DNA library (that is, growing more copies of the clones) is not as simple and straightforward as implied in the introduction. The range of insert sizes and the variety of DNA sequences in libraries makes nearly every clone unique; therefore, each host cell has a somewhat different task to perform in replicating its cloned insert. The inevitable consequence is that some host cells grow faster than others, and if libraries are amplified by simply growing more cells, some sequences will be over- or under-represented in the amplified library. This problem is relatively mild in phage libraries, but even they can become distorted in representation if they are amplified too many times. Nonrecombinant phage usually reproduce faster than recombinants, so they rapidly become the most common constituent in an overamplified library. Therefore, phage libraries are best handled by amplifying them only once, by one seeding on a lawn of *E. coli* cells, and freezing aliquots of the harvested phage particles for future use or for sharing with other laboratories.

In the case of cosmid libraries, it is usually disastrous to grow the clones in close proximity to and therefore in competition with one another. We found that a single amplification of a cosmid library in which 2 to 3 percent of the clones were nonrecombinant produces a library consisting of 40 percent nonrecombinants. Such a result is clearly unacceptable, especially for libraries that are difficult to construct. Unfortunately, the solutions to this problem are labor-intensive. Perhaps the best method is to lightly seed a primary library on agar plates, allow each bacterium to form a small colony, and transfer a portion of each colony with a toothpick into a well in a 96-well microtiter plate. A part of each colony can then be moved to another microtiter plate with a 96-prong hand stamp or by a robot, if one is available. This procedure ensures that each colony will survive, but for a library including the

Libraries from Flow-sorted Chromosomes

Larry L. Deaven

For any kind of genomic-DNA library, subdividing the DNA of the entire genome before library construction is almost always advantageous. The resulting set of libraries includes all of the genomic DNA, but each library is less complex than a single library containing all of the cellular DNA. A natural way to make subsets of human DNA is to make a separate library for each chromosome. To include all of the nuclear DNA in human cells, 24 different libraries are necessary (22 autosomes plus the X and Y chromosomes). The libraries vary in size, the largest (for chromosome 1) being five times as large as the smallest (for chromosome 21).

The most efficient way to make chromosome-specific libraries is to start with flow-sorted chromosomes. Los Alamos scientists pioneered the technology of flow sorting chromosomes as a direct result of the invention and development of flow cytometers at the Laboratory during the 1970s. Figure 1 diagrams flow sorting as we use it in making DNA libraries.

The first libraries made from sorted chromosomes at Los Alamos were from Chinese hamster chromosomes. Those chromosomes are larger and better differentiated from one another by base-pair content than are human chromosomes, properties which make them relatively easy to sort on a flow cytometer. On the basis of that success, we thought it would be feasible to construct certain types of libraries from sorted human chromosomes. The Department of Energy agreed to support the work, and because the scope of the envisioned project was large, we asked our colleagues at Lawrence Livermore National Laboratory if they would join in an effort to make a complete set of chromosome-specific libraries. Our initial discussions in 1983 led to the National Laboratory Gene Library Project, which continues to be a component of the Human Genome Centers at the two laboratories.

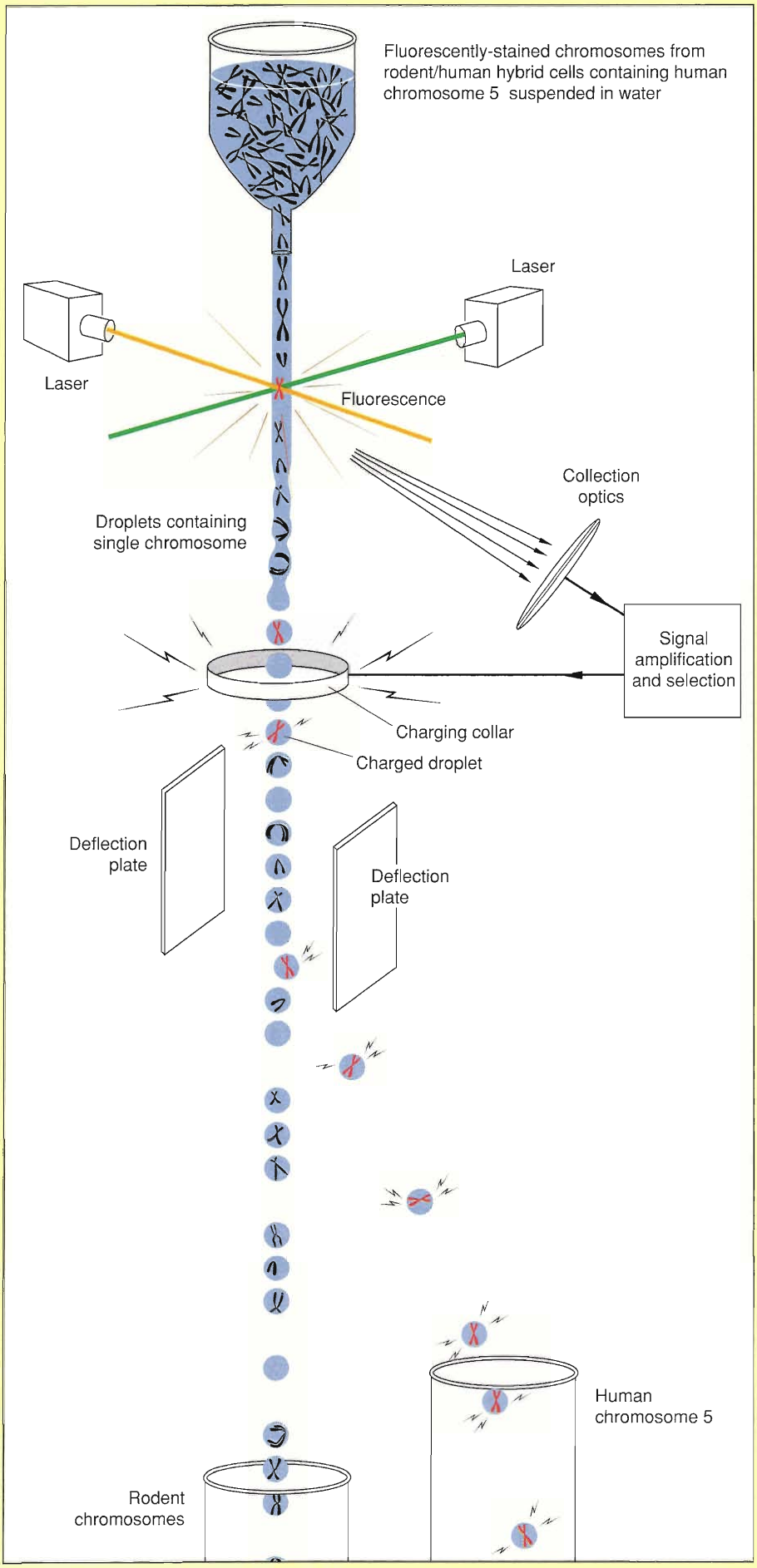
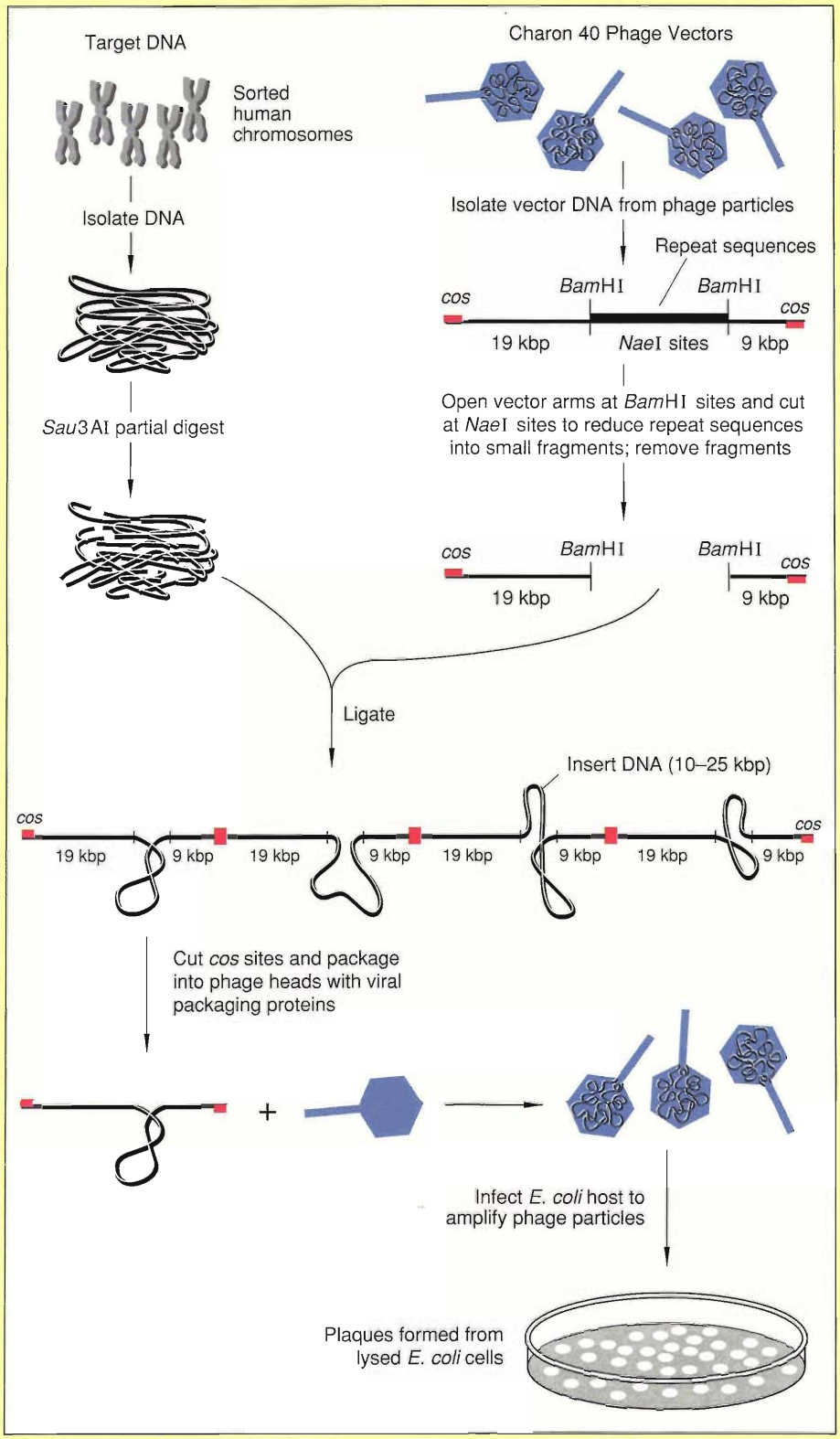


Figure 1. Purifying Chromosomes through Flow Sorting

Flow sorting provides a way of separating chromosomes of one type from a mixture. The example in the illustration is the separation of human chromosome 5 from rodent chromosomes all isolated from a rodent/human hybrid cell line. A liquid suspension of metaphase chromosomes is carried through the flow sorter in a narrow stream. The chromosomes have been stained with two fluorescent dyes: Hoechst 33258, which binds preferentially to AT-rich DNA, and chromomycin A_3, which binds preferentially to GC-rich DNA. The stained chromosomes pass through a point on which two laser beams are focused, one beam to excite the fluorescence of each dye. Each chromosome type has characteristic numbers of AT and GC base pairs, so chromosomes can be identified by the intensities of the fluorescence emissions from the two dyes. If the fluorescence intensities indicate that the chromosome illuminated by the lasers is the one desired, the charging collar puts an electric charge on the stream shortly before it breaks into droplets. When droplets containing the desired chromosome pass between charged deflection plates, they are deflected into a collection vessel. Uncharged droplets lacking the desired chromosome go into a waste-collection vessel. The flow instruments used at Los Alamos can analyze 1000 to 2000 chromosomes per second and sort approximately 50 chromosomes per second.

Figure 2. Phage Cloning Using Sorted Human Chromosomes as Target DNA

The phage vector (Charon 40) used to construct libraries from flow-sorted human chromosomes at Los Alamos contains a *cos* site, a large number of restriction sites, and a removable section consisting of repeat sequences (see Figure 6 in the main text). When the vector is used for cloning, the section of repeat DNA is cut into small pieces and discarded. The removal provides space for insert DNA. The vector consists of a 19-kbp arm and a 9-kbp arm, leaving room for inserts of 10 to 25 kpb. After the vector DNA has been isolated from phage particles, it is digested with the restriction enzymes *Bam*HI and *Nae*I. The eightyfold-repeated sequence constituting the central portion of Charon 40 contains an *Nae*I site, so the central portion is cut into small pieces by the *Nae*I digestion. The *Bam*HI digestion provides cloning sites on one end of each vector arm. Because *Bam*HI and *Sau*3AI produce identical sticky ends, the cloning sites are compatible with the *Sau*3AI sites on the ends of each fragment of partially digested target DNA. When the vector arms are ligated with fragments of target DNA, a concatamer forms that is cut at the *cos* sites to form individual recombinant phage chromosomes. These chromosomes are packaged into phage particles which then infect *E. coli* cells.



In 1983 the Human Genome Project did not exist. It was too early to seriously consider the construction of a physical map and the sequencing of the entire genome. Genetic mapping, on the other hand, was enjoying a period of unprecedented growth. The theory and methodology of finding genes using DNA markers had been developed, and major efforts were under way to locate human disease genes and to develop high-resolution genetic maps (see "Modern Linkage Mapping" in "Mapping our Genes"). Accordingly our first aim for the library project was to construct a phage library of small DNA inserts for each human chromosome. Small inserts were desirable for two reasons. First, the major challenge in making libraries from sorted chromosomes is to maximize the efficiency of each step in the cloning procedure in order to be able to make large libraries from small amounts of sorted DNA. In 1983, the technology for making small-insert (complete-digest) libraries was more reliable and could start with smaller amounts of target DNA than that for large-insert (partial-digest) libraries, which require cosmids. The second reason was the utility of small-insert libraries to genetic mappers. Repetitive DNA sequences are dispersed throughout the human genome, and the larger the insert, the more likely it is to contain at least one sequence repeated elsewhere. Probes containing repetitive sequences hybridize to many sites in the genome unless the repeat sequence is blocked. Single-copy probes identify only one site, a useful step in genetic mapping.

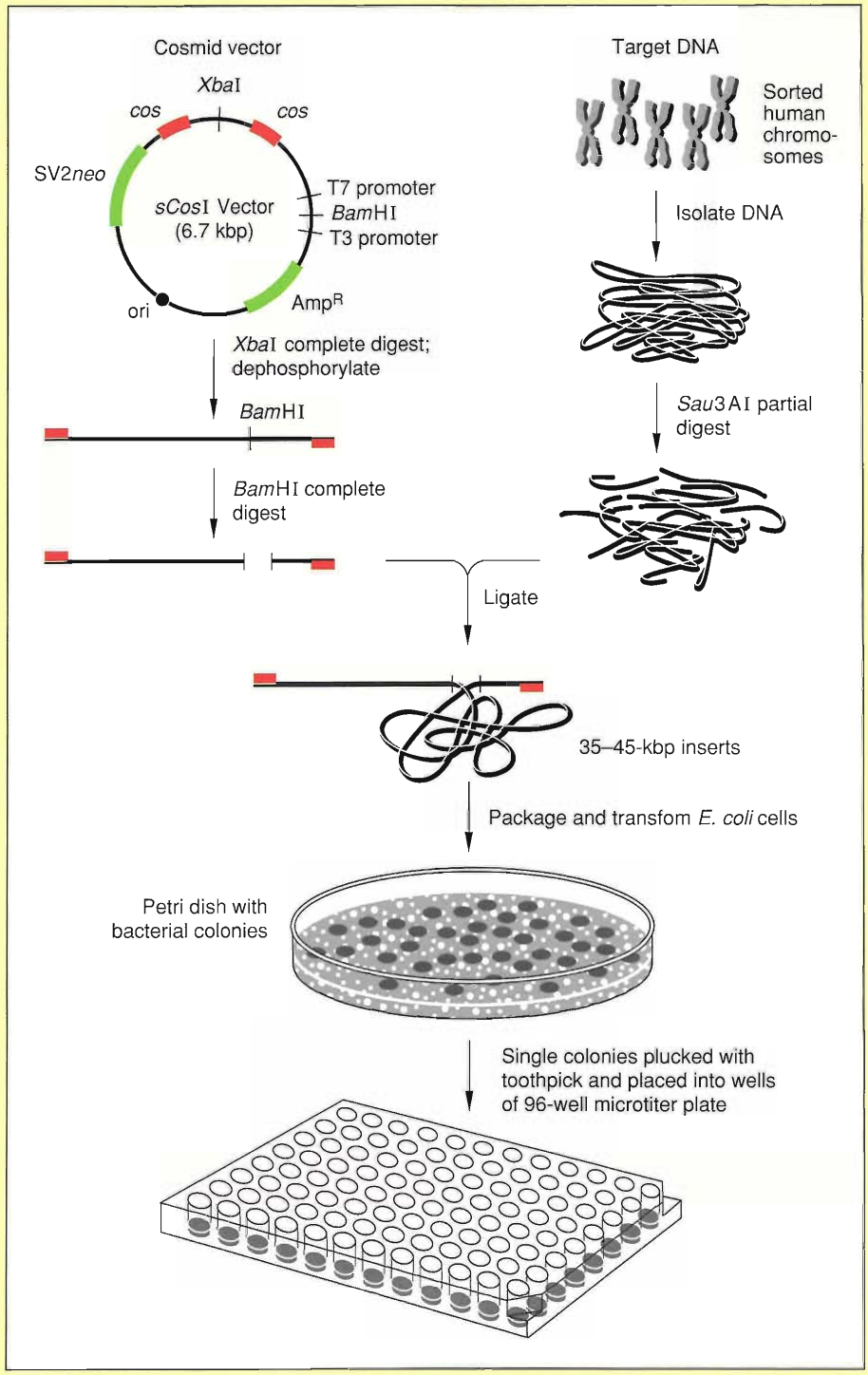
Our strategy for the first set of libraries made from sorted chromosomes was to digest the chromosomal DNA completely with a six-base cutter and to clone the fragments into a λ-phage vector called Charon 21A. Such a restriction enzyme reduces DNA to fragments having an average length of 4 kbp. However, approximately a third of the DNA is in fragments larger than 9 kbp, the upper limit for acceptance by Charon 21A. To reduce the amount of uncloned DNA, we constructed for each chromosome two libraries using different restriction enzymes; the Los Alamos project used *Eco*RI, while the Livermore project used *Hin*dIII. We estimate that at least 90 percent of the chromosomal DNA is contained in the two libraries together.

Our small-insert libraries were amplified one time, then sent to the American Type Culture Collection in Rockville, Maryland, where they are stored in liquid nitrogen. Samples from the original libraries are available to research groups throughout the world. They have been used extensively as a source of probes for polymorphic markers used in mapping genes, especially genes that can cause diseases. For example, as part of the searches for the defects responsible for cystic fibrosis and Huntington's disease, several hundred probes have been isolated from the chromosome-4 and chromosome-7 libraries and mapped to those chromosomes. Although improved methods now permit the construction of larger-insert libraries, the Los Alamos and Livermore complete-digest libraries are still useful. Over 4000 samples have been sent to research laboratories.

As we were finishing construction of the complete-digest libraries, it became obvious that chromosome-specific libraries with larger inserts were highly desirable. For molecular studies of gene structure and expression, they would have the advantage of

Figure 3. Cosmid Cloning of DNA from Sorted Human Chromosomes

The cosmid vector (sCos 1) contains two *cos* sites for rejoining the linear recombinant molecule after transformation. It also contains two selectable markers [resistance to ampicillin (amp^R) and to neomycin (SV2*neo*)], a number of restriction sites, a plasmid replicon including an origin of replication (ori), and promoter sequences from the T3 and T7 phage. The T3 and T7 promoters are used to generate end probes, as discussed in the section on YACs in the main article. The vector molecule is linearized by cutting with the restriction enzyme *Xba*I, then separated into two cloning arms by cutting with *Bam*HI. After fragments between 35 and 45 kbp in length are ligated to the vector arms, the recombinant DNA molecules thus produced are packaged into phage protein coats. The resulting infectious phage particles insert the recombinant molecules into *E. coli* cells where the molecules cyclize and live as plasmids. To prevent the faster-growing *E. coli* cells from overwhelming the slower ones, each colony is placed in a separate well of a microtiter plate.



containing whole genes or even groups of genes in a single cloned insert. Moreover, molecular biologists were then discussing and planning the mapping and sequencing of the entire human genome. Large-insert libraries from each chromosome would be a valuable resource for those massive tasks. The entire human genome in a cosmid library can be thought of as a jigsaw puzzle of 75,000 pieces; the chromosome-specific libraries would be 24 puzzles with an average of 3125 pieces in each.

During the years we spent constructing small-insert libraries, significant improvements were made in the efficiency of vector systems capable of carrying large inserts. The most important improvement for our large-insert project was the construction of cosmid vectors with two *cos* sites instead of one. Such cosmid vectors can be cleaved into two cloning arms, each with a *cos* site at one end. The cosmid arms can then be ligated to the partially digested human target-DNA fragments, much as in phage cloning. Each resulting recombinant molecule consists of two cloning arms each ligated to an end of a fragment of human DNA. If the *cos* sites are between 30 kbp and 52 kbp apart, the recombinant molecule can be packaged in vitro to produce infectious phage particles. Using this cloning system, a cosmid library with inserts 35 to 45 kbp in length can be made from less than a microgram of DNA.

The laboratories' joint strategy for the construction of a second set of libraries with larger inserts was to divide the human chromosomes between Los Alamos and Livermore. Each laboratory would construct a partial-digest phage and cosmid library for the chromosomes assigned to it. Los Alamos has made libraries for chromosomes 4, 5, 6, 8, 11, 13, 16, and 17; Livermore, for chromosomes 19, 22, and Y.

Our current work incorporates several changes in the construction and handling of libraries. All chromosomes are sorted from hybrid-cell lines rather than from human cells because of the advantages discussed in the main text. The phage libraries, illustrated in Figure 2, have inserts 10 to 25 kbp long. They are stored as pools of clones in a liquid medium and distributed as samples like the small-insert libraries.

As illustrated in Figure 3, the cosmid libraries are seeded on Petri plates. The libraries are then arrayed, that is, single colonies are transferred to 96-well microtiter plates. Enough colonies are isolated to cover the chromosome five times. A chromosome of average size requires about 5×3125 or 15,625 colonies. The inserts have not yet been characterized, so we do not know whether the DNA in the inserts covers the entire chromosome. When all the colonies have been transferred, we make five to ten copies of each microtiter plate. Sets of microtiter plates are sent to laboratories involved in projects to map the entire chromosome or a major portion of it. In addition, the colonies in one set of plates are allowed to grow to high density, and then the bacteria are removed from each well and pooled. Laboratories interested in isolating one or a few genes on the chromosome can obtain portions of the pooled library.

An advantage of storing a library in a set of microtiter plates is that each clone has an alphanumerically labeled location. The labeling permits all the data on the

clones from different laboratories to be combined for analysis. Ideally, all interested laboratories should have copies of the plates; however, distribution of so many copies would be too expensive.

The partial-digest libraries that have been completed are major resources for laboratories constructing physical maps for chromosomes 4, 5, 8, 11, 16, 17, and 19. The libraries are used directly in assembling contigs of cosmids and also contribute to physical mapping with YACs. In order to make a high-resolution map from YAC contigs, each YAC must be subcloned into cosmid or phage vectors, a time-consuming process. A more rapid way to find cosmids that are part of a YAC is to screen an arrayed cosmid library with DNA from the YAC insert. A second major use of the partial-digest libraries is in the isolation of genes for detailed studies of normal and abnormal structure and expression. A third use is the identification of specific chromosomes or parts of chromosomes. Each library is very pure, and the inserts in it can be labeled with fluorescent stains and hybridized in situ to cells or metaphase chromosomes. In interphase cells hybridization reveals the nuclear location of the chromosome represented in the library. In metaphase chromosomes hybridization identifies only the pair of chromosomes that the library represents. If a piece of a labeled chromosome has been broken and has translocated to another chromosome, the translocation is easily visible. The latter application is revolutionizing the detection of chromosomal rearrangements induced by substances that break chromosomes and by diseases like cancer, in which rearranged chromosomes are common.

Although our cosmid libraries are not yet complete, during the past two years we have devoted a substantial portion of our library-construction effort to YAC cloning. We were fortunate in having Mary Kay McCormick join our Center in 1989. Before coming to Los Alamos, she had demonstrated the feasibility of using sorted chromosomes as the source of target DNA in making YACs. To construct a YAC library, we had to overcome two major obstacles. Long pieces of human DNA had to be obtained from sorted chromosomes, and YAC-cloning techniques had to be optimized in order to use the small amounts of DNA available after sorting. Solutions to both problems were found through the skills of dedicated investigators. Chromosome isolation and flow sorting must be accomplished without delay because DNA degradation begins as soon as the chromosomes are extracted from the cells. To sort 1-microgram samples of DNA in a limited time, sorting continues around the clock. The sorted chromosomes are collected in agarose plugs which hold the DNA in the stable agarose matrix and protect it from shear stresses during isolation from the chromosome and digestion with restriction enzymes. The agarose is then melted so that the vector arms and DNA ligase can be mixed in. After ligation the recombinant molecules are fractionated by preparative pulsed-field gel electrophoresis, which concentrates all the DNA fragments longer than 200 kpb into a single band in the gel.

To facilitate transformation, the walls of yeast cells are removed. (Yeast cells without walls are called spheroplasts.) The long recombinant DNA molecules are added to

the spheroplasts in the presence of the polyamines spermine and spermidine, which are believed to bind to and condense DNA. To obtain large numbers of recombinant yeast colonies (as many as 2400 have been obtained from 1 microgram of target DNA), all of the above steps must work well. Probably the most frustrating step is transforming the yeast cells. It is difficult to control, it sometimes fails, and because it is the last cloning step, failure means that all the previous work must be repeated.

We have completed two YAC libraries, one for chromosome 16 and one for chromosome 21. Both libraries were made from target DNA completely digested with restriction enzymes that have infrequent cleavage sites. Therefore, how completely the libraries represent chromosomal DNA depends on how uniformly the cleavage sites are distributed along the chromosomes. We will not know the completeness of the representation until we have generated a considerable amount of data on each library. Preliminary results suggest that the YACs made from digests with *Eag*I or with a combination of *Not*I and *Nhe* are clustered near certain chromosomal regions such as the centromere, but that YACs made from *Cla*I digests may be more uniformly distributed. We are attempting to ensure that future YAC libraries have unbiased distributions by making them from partial digests. Other studies of the libraries suggest that the frequency of chimeric inserts is quite low. Fifty-three YAC inserts have been hybridized in situ to chromosome 21. None of them hybridized to more than one region of the chromosome, which would have been evidence of a chimera.

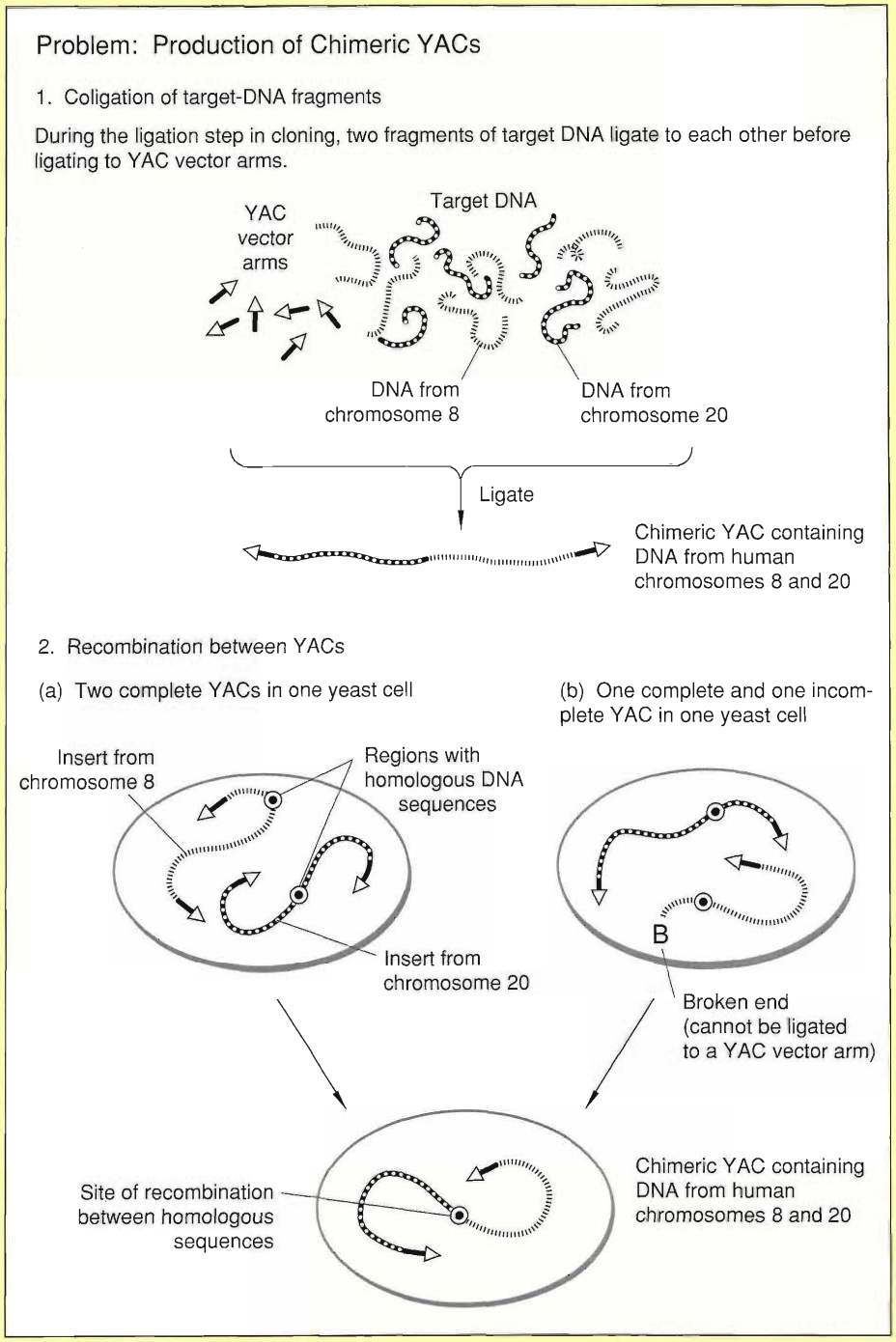
The reasons for the absence of chimeric inserts are not completely clear. We took a number of steps intended to reduce their frequency. As illustrated in Figure 4, chimeric YACs are believed to originate either from ligation of two pieces of target DNA or from recombination between two YACs after they have both transformed the same yeast cell, especially when at least one YAC is incomplete.

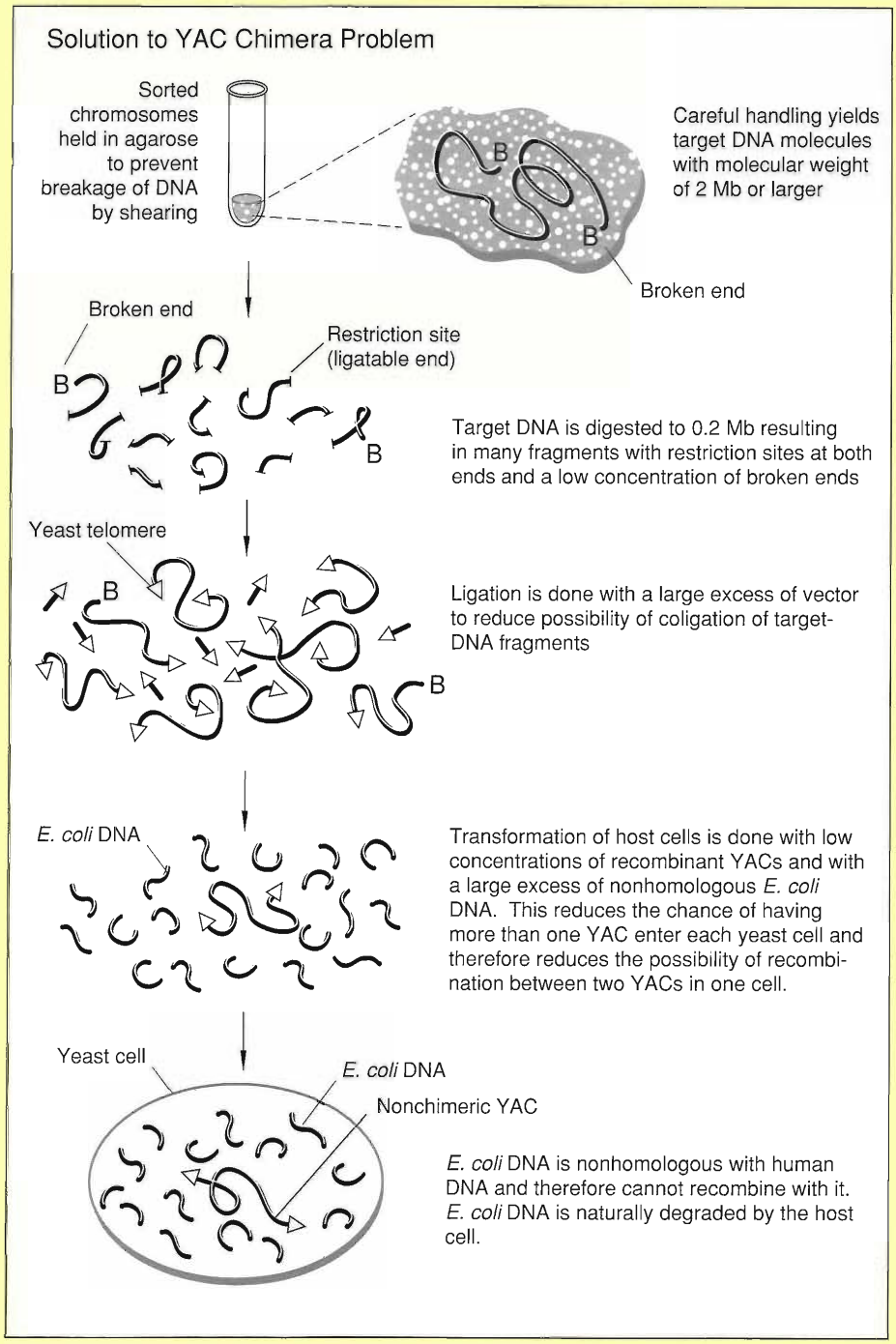
To minimize the coligation of target DNA, we added much more vector DNA to the ligation mixture than the restriction fragments could react with. To reduce the possibility of recombination inside yeast cells, we took two precautions. The first was to handle target-DNA restriction fragments so as to minimize breakage. The second was to attempt to limit the possibility that more than one YAC would enter a single spheroplast by diluting the YACs to the point where it was unlikely that two YACs would enter the same spheroplast.

Although sufficient data are not yet available to thoroughly evaluate the chromosome-specific YAC libraries, all evidence to date suggests that they will be a valuable resource for constructing physical maps of chromosomes. The libraries combine the advantages of large insert size and division into subsets of the genome to provide the least complex mapping elements available. They are being used to close the gaps between cosmid contigs in the Los Alamos chromosome-16 map, and they should prove to be excellent sources of fragments for the initiation of maps of other chromosomes. We expect the Library Project now to focus on the construction of large-insert libraries in YACs and other cloning systems under development.

Figure 4. Chimeric YACs

Part A shows two causes of chimeric YACs. The first is that, since target DNAs are all cut with the same restriction enzyme, they can ligate to each other. The resulting chimeric insert can then ligate to vector arms. The second is that if two YACs enter the same yeast cell and their inserts have homologous sequences, they can recombine with each other, producing a chimeric YAC. Recombination is especially likely if one or both of the YACs is incomplete, either because the insert is broken or because it ligated to only one vector arm. Part B shows our solutions to the problem. To limit breaking we keep sorted chromosomes in agarose and handle the DNA carefully. Our target DNA molecules are typically longer than 2000 kbp. Since the restriction digest produces fragments averaging 200 kpb, few fragments have broken ends. Then we add many more vector molecules than insert molecules to the ligation reaction, making ligation between two insert molecules unlikely. During transformation, to reduce the probability that two YACs enter a yeast cell, we add *E. coli* DNA, which is not homologous with human DNA. That step greatly dilutes the YACs while keeping the total DNA concentration high enough to induce transformation.




Solution to YAC Chimera Problem
Sorted chromosomes held in agarose to prevent breakage of DNA by shearing
B
B
Careful handling yields target DNA molecules with molecular weight of 2 Mb or larger
Broken end
Broken end
Restriction site (ligatable end)
B
B
Target DNA is digested to 0.2 Mb resulting in many fragments with restriction sites at both ends and a low concentration of broken ends
Yeast telomere
B
B
Ligation is done with a large excess of vector to reduce possibility of coligation of target-DNA fragments
E. coli DNA
Transformation of host cells is done with low concentrations of recombinant YACs and with a large excess of nonhomologous E. coli DNA. This reduces the chance of having more than one YAC enter each yeast cell and therefore reduces the possibility of recombination between two YACs in one cell.
Yeast cell
E. coli DNA
Nonchimeric YAC
E. coli DNA is nonhomologous with human DNA and therefore cannot recombine with it. E. coli DNA is naturally degraded by the host cell.

Library Distribution

The success of the people working in the Library Project has created a need for large-scale duplication of clones in microtiter plates. The Los Alamos portion of the cosmid-library project will require copying over 200,000 clones six to ten times, and our future work in YAC-library construction will produce more clones to be copied. As important as duplicating clones in microtiter plates is making replicas of microtiter plates as spots on nylon membranes, a procedure that provides a convenient way to screen an entire library. A 96-prong stamp is inserted into the wells of a microtiter plate and then gently placed on a membrane. The bacteria collected on each prong are transferred to the membrane. The membrane rests on an agar culture medium from which the bacteria absorb nutrients. The resulting 96 colonies in the form of spots on the membrane can then be screened with a DNA probe. Any spots that hybridize with the probe DNA can be identified and the corresponding clones can be located in the microtiter plate. Those clones can then be selected and regrown for further analysis. We use this screening procedure extensively in our construction of a map of chromosome 16, and we currently use it to share our libraries with other laboratories. For example, an investigator at the Institute of Cytology of the Russian Academy of Sciences is interested in finding inserts that come from a region of chromosome 5. We sent her a set of membranes containing spots from each microtiter well in the arrayed chromosome-5 library. She probed the membranes with her collection of probes from the region she was interested in, and we selected and shipped colonies corresponding to each of the spots that tested positive. Duplicating and shipping copies of the library in microtiter plates is expensive, and we hope that the use of membranes will prove to be a useful alternative.

To help us meet the demands of library duplication, a group of robotics engineers at Los Alamos has constructed a robot capable of accomplishing that task. The robot can choose a microtiter plate from a dispenser, scan the barcode label on the plate, and insert a 96-prong tool into the wells in the plate. The robot then presses the tool against a membrane, transferring spots of bacteria from the prongs to the membrane. Finally it sterilizes the tool, replaces the lid on the microtiter plate, and returns the plate to a stacker. The robot can transfer colonies to the same membrane up to 96 times, each time shifting the position of the tool slightly, and thus can vary spot densities from 576 to 9216 per 22-cm^2 membrane. The robot's versatility is valuable; because denser packing of spots is more efficient but may be harder to read, different densities are suitable for different applications. ■

Further Reading

L. Scott Cram, Dale M. Holm, and Paul F. Mullaney. 1980. Flow cytometry: A new tool for quantitative cell biology. *Los Alamos Science*, volume 1, number 1.

L. Scott Cram, Larry L. Deaven, Carl E. Hildebrand, Robert K. Moyzis, and Marvin Van Dilla. 1985. Genes by mail. *Los Alamos Science*, number 12.

entire human genome it requires picking 75,000 colonies by hand. Another procedure is to seed the primary library on a series of filter membranes laid on agar surfaces. After colonies form, the original filters can be copied by pressing them against additional filters. This method is somewhat less tedious than the one previously described, but colonies may be lost if they do not transfer and regrow from the master filters.

Yeast colonies must be handled individually and are usually placed in microtiter plates. Since only about 7500 YACs would be needed to cover the human genome, library distribution is much less labor-intensive than for cosmids.

All libraries can be stored indefinitely by freezing them at -70°C. Before the colonies are frozen, they are suspended in their growth medium supplemented with 30 to 40 percent glycerol; the glycerol protects cellular structures from damage by ice-crystal formation.

Problems and Errors in Cloning

The previous discussion of vectors may make cloning seem more straightforward than it is. All cloning systems involve difficulties, especially the newer ones that have not had the benefit of years of testing and improvement. It would be unfair to the people who diligently and carefully perform this work not to describe some of the pitfalls that can be encountered.

A problem common to all cloning systems that has not yet been discussed in detail is the occurrence of unwanted ligations when vector and target DNAs are joined with DNA ligase. Undesirable ligations include the religation of the ends of a linearized plasmid and the joining of two phage or YAC arms. In many cases, such ligations would result in nonrecombinant contaminants of a library, which in some cases would be indistinguishable from recombinants. Another undesirable process is the ligation of two small fragments of target DNA, which may later be cloned as a chimeric insert. The standard practice to avoid these ligations is to treat either the vector or the target DNA with an enzyme called calf intestinal alkaline phosphatase (CIP). This enzyme removes phosphate groups from the 5′ ends of linear DNA. Because DNA ligase cannot join DNA molecules unless the 5′ ends have phosphate groups, undesirable ligations between treated molecules can not happen.

With some phage and cosmid vectors, ligation between vector molecules does not cause problems because the vector DNA is not large enough for proper packaging and therefore a vector religation does not result in a viable nonrecombinant. In those cases the target DNA rather than the vector DNA is treated with CIP. That method is very useful in preventing the formation of chimeric inserts, especially when the target DNA contains small fragments.

Unfortunately, all the CIP must be removed to make the subsequent ligations work efficiently. CIP is removed by digestion with another enzyme called proteinase K, followed by extraction of protein-degradation products with phenol and chloroform. Since those steps require handling the DNA, they increase the risks of shearing and of degradation by nonspecific nucleases (DNA-digesting enzymes that are common contaminants in biochemicals). Therefore CIP treatment is seldom used for the large and consequently fragile fragments needed for YAC constructions. The protocols for CIP treatment must also be carefully controlled because an incomplete treatment would result in a library of questionable value. Furthermore, a batch of CIP that contains nucleases can destroy painstakingly prepared target DNA.

An example of the tricky nature of library construction is the loss of restriction-enzyme specificity, a phenomenon called star activity. Earlier in this article restriction enzymes were described as being specific for one DNA sequence. For some enzymes, this is not completely true. Their specificity may be altered when they are used under altered reaction conditions. These altered conditions include high enzyme concentration, use of manganese instead of magnesium, low concentration of electrolytes, high pH, or the presence of organic solvents such as glycerol. If DNA is digested with *Eco*RI, for instance, under any of these conditions, the enzyme can cleave DNA at sequences that differ from the normal recognition sequence by a one-base substitution. The result of star activity is a library some of whose clones are jumbles of small pieces of vector and insert DNA.

New Directions in Library Construction

Libraries for Constructing STS Markers. An STS library is a chromosome-specific library designed to facilitate identification and cloning of STSs (sequence-tagged sites) from one human chromosome (see "The Polymerase Chain Reaction and Sequence-tagged Sites" in "Mapping the Genome"). The inserts in an STS library are cloned in M13 vectors. Since the M13 cloning system is efficient, STS libraries can be made with very small amounts of sorted DNA or from the DNA in a rodent-human hybrid cell line containing a single human chromosome. The target DNA is digested to completion with one or two frequent cutters, and then it is ligated with M13 double-stranded DNA. The resulting libraries of M13 clones

have small (200 to 1000 base pairs) inserts that are in a useful form for the dideoxy chain-termination sequencing method.

After each cloned insert has been sequenced, the sequence is searched for a single-copy sequence of 200 to 300 base pairs that can be used as an STS. The polymerase chain reaction can then be used to locate the clones in a total genomic YAC library that contain the identified STSs. In this approach, a small amount of sorted DNA is used to make an M13-based chromosome-specific library that can provide hundreds of STS markers for each chromosome.

Microdissection Libraries. Microdissection libraries are made from a specific region of a chromosome and are usually very small libraries, perhaps containing only a few inserts. The target DNA may be from a single chromosome band or from an area containing a defect such as a visible gap or fragile site. Target DNA may be obtained by fixing a chromosome on a microscope slide and scraping off and collecting an identifiable region. An alternative method is to use a laser to burn away all of the chromosome except the region of interest. The tiny amounts of DNA obtained are usually amplified using PCR and then cloned into a phage vector that accepts small inserts. These libraries are useful as probes to determine which cosmids or YACs from other libraries contain inserts that cover the dissected region. Probes from microdissection libraries have been used effectively to screen chromosome-specific libraries constructed from flow-sorted chromosomes.

cDNA Libraries. The synthesis of a cDNA probe for the human β-globin gene as early as 1975 was made possible by a unique feature of reticulocytes, the precursors of red blood cells. Reticulocytes produce large amounts of hemoglobin and contain very little mRNA other than the globin mRNAs. Therefore the mRNA extracted from human reticulocytes is essentially pure globin mRNAs. Once extracted, the globin mRNAs are reverse transcribed (by the enzyme reverse transcriptase) into cDNAs that hybridize to those clones in a human genomic library that contain all or a portion of each of the human globin genes. Similarly, mRNA extracted from cells of the pituitary gland has been used to isolate the growth-hormone gene.

The abundance of one or a few mRNAs in certain specialized cells makes synthesizing cDNA probes for the corresponding genes relatively easy. However, in most cells some 10,000 genes are expressed at different levels, and the copy numbers of the corresponding mRNAs range from 1 to 20,000. To facilitate screening a library of the cDNAs synthesized from such a population of mRNAs, the cDNAs are cloned in special plasmid or λ-phage vectors in which the cloning site is embedded within the bacterial gene for β-galactosidase. The host bacterial cell "recognizes" the β-galactosidase gene and transcribes not only the β-galactosidase gene but also the foreign cDNA insert. If the insert is in the right orientation and in the same reading frame as the bacterial gene, the result is a fusion protein consisting of part of β-galactosidase attached to part of the polypeptide product of the mRNA. A labeled antibody to the protein product corresponding to a cDNA of interest can then be used to select the clone or clones containing the cDNA of interest. (An antibody to a certain protein binds only to that protein.)

To reduce the labor involved in screening a cDNA library, attempts have been made to reduce the number of different cDNAs present in the target DNA by preparing the target DNA from the mRNAs that are present in one cell type but not in another. The mRNA from cell type 1 is reverse transcribed into single-stranded cDNA, which is then allowed to hybridize with a larger quantity of the mRNA from cell type 2. The cDNA that remains single-stranded corresponds to the mRNA that is present only in cell type 1. A library made from that cDNA contains fewer cDNA species and is therefore easier to screen than a library of the cDNAs corresponding to all the mRNAs present in cell type 1.

More recently attempts have also been made to construct normalized, or equalized, cDNA libraries. The ideal normalized cDNA library would not only be normalized (contain an equal number of clones of each cDNA) but would also be complete (contain all the cDNAs corresponding to all the mRNAs present in any cell of the organism at any time during its life). No complete normalized cDNA library is yet available, but cDNA libraries that are close to being normalized are available for certain human tissues. The procedure for normalizing libraries begins with the synthesis of the cDNAs corresponding to all the mRNAs in a selected tissue and cloning the cDNAs in λ-phage vectors. The cloned inserts are amplified by PCR, denatured, and allowed to renature. Because the abundant cDNA species renature more rapidly than the rare species, the abundances of the cDNA species that remain single-stranded vary by a factor much smaller than the original 20,000. In fact, variation by a factor of 40 has been achieved. An obvious application of a normalized cDNA library is as a source of probes for selecting clones from other libraries and locating genes on physical maps.

The continuing need for reliable and efficient cloning systems capable of propagating inserts larger than 45 kbp

(the upper limit for cosmids), has led to the development of several alternatives to YACs, all of which are still being improved. The perfect cloning system for a library, by today's standards, would accept inserts in the range of 200 to 300 kbp. With inserts of that size a library would not need an excessive number of clones to cover the human genome and would still allow genes to be located with a useful degree of precision. The ideal system would have all the features mentioned in the discussion of host cells earlier, particularly low frequencies of chimera formation, clone loss, and deletion of inserts (which are the major disadvantages of YACs). In addition, all human sequences should be clonable in the system, so that libraries can cover the entire genome and any desired region can be located by using an STS.

Cloning systems have evolved steadily since the 1970s and new types of libraries will continue to be developed as new applications arise. The Laboratory has pioneered the construction of chromosome-specific libraries (see "Libraries from Flow-sorted Chromosomes"). That work too is evolving in response to the challenges presented by the Human Genome Project and by the rapid progress of molecular genetics. ■

Further Reading

Stanley N. Cohen. The manipulation of genes. *Scientific American*, July 1975, 25–33.

L. L. Deaven, M. A. Van Dilla, M. F. Bartholdi, A. V. Carrano, L. S. Cram, J. C. Fuscoe, J. W. Gray, C. E. Hildebrand, R. K. Moyzis, and J. Perlman. 1986. Construction of human chromosome-specific DNA libraries from flow-sorted chromosomes. Cold Spring Harbor Symposia on Quantitative Biology 51:159.

Ernst-L. Winnacker. 1987. *From Genes to Clones: Introduction to Gene Technology*, translated by Horst Ibelgaufts. New York: VCH Publishers.

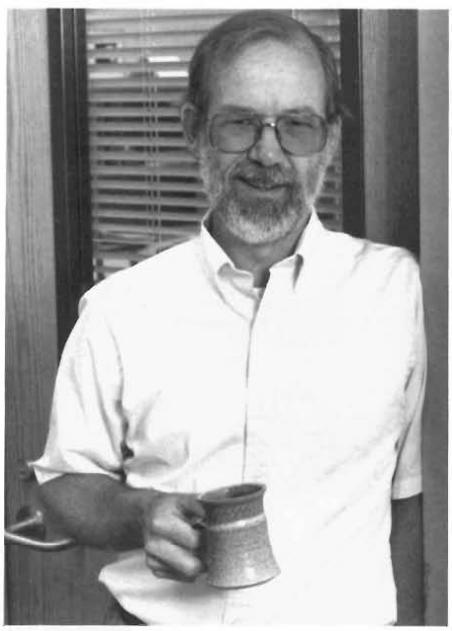
Shelby L. Berger and Alan R. Kimmel, editors. 1988. *Guide to Molecular Cloning Techniques*. Methods in Enzymology, volume 152. San Diego: Academic Press.

J. Sambrook, E. F. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*, second edition. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

David A. Micklos and Greg A. Freyer. 1990. *DNA Science: A First Course in Recombinant DNA Technology*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Larry L. Deaven. 1991. *Chromosome-Specific Human Gene Libraries*. In *Encyclopedia of Human Biology*, volume 2, Renato Dulbecco, editor-in-chief. San Diego: Academic Press.

James D. Watson, Michael Gilman, Jan Witkowski, and Mark Zoller. 1992. *Recombinant DNA*, second edition. New York: W. H. Freeman and Company.



Larry L. Deaven is the principal investigator of the National Laboratory Gene Library Project at Los Alamos and Deputy Director of the Los Alamos Human Genome Center. [See "Members of the Human Genome Center at Los Alamos National Laboratory" for biographical details.]